

Mascot:

Installation and Setup

©2007 Matrix Science Ltd. All rights reserved.

The information contained in this publication is for reference purposes only and is subject to change at any time. Every effort has been made to supply complete and accurate information. However, Matrix Science Ltd. assumes no responsibility and will not be liable for any consequential loss or damages that might result from the use of this manual or from any errors or omissions in the information contained herein.

No part of this document may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of Matrix Science Ltd.

Mascot is a trade mark of Matrix Science Ltd. All third party trade marks and service marks referred to in this publication are hereby acknowledged.

Matrix Science Limited
64 Baker Street
London W1U 7GB
UK

Phone +44 (0)20 7486 1050

Fax +44 (0)20 7224 1344

Email info@matrixscience.com

WWW <http://www.matrixscience.com/>

February 2007: Revision 2.2.0

The latest version of this publication can be found on the Mascot distribution CD-ROM as an Adobe Acrobat document, filename *manual.pdf*

End User Licence Agreements

MASCOT PROTEIN IDENTIFICATION SYSTEM

END USER LICENCE AGREEMENT

IMPORTANT – PLEASE READ CAREFULLY: This End User Licence Agreement is a legally binding contract between you (either an individual or a single corporate entity) and Matrix Science Limited for the product identified above, which includes computer software, electronic documentation, printed documentation, and any subsequent updates and supplements (the “Software”).

By installing or using the Software, you agree to be bound by the terms of this agreement. If you do not agree to the terms of this agreement, we are unwilling to license the Software to you. In this case, do not install or use the Software. Return the Software to Matrix Science Limited or their authorised distributor within 30 days of receipt for a full refund.

1 Licence

Matrix Science Limited owns the copyright in the Software contained within this package and all other copies which you are authorised by this agreement to make.

This licence is personal to you (either an individual or a single corporate entity) as the purchaser of a licence to use the Software and the licence granted herein is for your benefit only.

You may not use the Software in any way that permits unlicensed access to the Software. In particular, individuals who are not party to this licence or the general public must not be permitted access to the Software through a public network such as the Internet.

2 Permitted Users

As purchaser of a licence to use the Software, you may, subject to the following conditions:

- 2.1 load the Software onto and use it on a single computer (of the type identified on the package) which is under your control; and

- 2.2 copy the Software for backup and archival purposes and make up to two copies of the documentation (if any) accompanying the Software provided that the original and each copy is kept in your possession and that your installation and use of the Software does not exceed that allowed by this agreement.
- 2.3 modify the HTML and Perl documents for sole use by yourself in connection with the Software.

3 Restrictions of Use

You may not:

- 3.1 load the Software into two or more computers at the same time. If you wish to transfer the Software from one computer to another, you must erase the Software from the first system before you install it onto a second system;
- 3.2 sub-license, assign, rent, lease or transfer the licence or the Software or make or distribute copies of the Software;
- 3.3 translate, reverse engineer, decompile, disassemble, modify or create derivative works based on the Software except as permitted by Law;
- 3.4 make copies of the Software except for backup or archival purposes as permitted hereunder;
- 3.5 use any backup copy of the Software (or allow anyone else to use such copies) for any purpose other than to replace the original copy in the event it is destroyed or becomes defective;
- 3.6 distribute copies of modified HTML or Perl documents; or
- 3.7 copy the written materials (except as provided by this agreement) accompanying the Software.

4 Title

As licensee, you own only the medium on which the Software is recorded. We shall at all times retain ownership of the Software.

5 Warranty

We warrant that for a period of ninety days (90) from the date of delivery ('the Warranty Period'):

- 5.1 the medium on which the Software is recorded will be free from defects in materials and workmanship under normal use. If the medium fails to conform to this warranty, you may, as your sole and exclusive remedy, obtain (at your option) either a replacement free of charge or a full refund if you return the defective medium to us or to your supplier during the Warranty Period; and

5.2 the copy of the Software in this package will materially conform to the documentation that accompanies the Software. If the Software fails to operate in accordance with this warranty, you may, as your sole and exclusive remedy, return all of the Software and the documentation to us or to your supplier during the Warranty Period, specifying the problem, and we will provide you either with a new version of the Software or a full refund (at your option).

6 Disclaimer

We do not warrant that this Software will meet your requirements or that its operation will be uninterrupted or error free. We exclude and hereby expressly disclaim all express and implied warranties or conditions not stated herein, so far as such exclusion is or disclaimer is permitted under the applicable law. THIS AGREEMENT DOES NOT AFFECT YOUR STATUTORY RIGHTS.

7 Liability

7.1 Our liability to you for any losses shall not exceed the amount you originally paid for the Software.

7.2 In no event will we be liable to you for any indirect or consequential damages even if we have been advised of the possibility of such damages. In particular, we accept no liability for any programs or data made or stored with the Software nor for the costs of recovering or replacing such program or data.

7.3 Nothing in this clause limits our liability to you in the event of death or personal injury resulting from our negligence.

8 Termination

8.1 The agreement and the licence hereby granted to use the Software automatically terminates if you:

8.1.1 fail to comply with any provisions of this agreement; or

8.1.2 voluntarily return the Software to us.

8.2 In the event of termination in accordance with clause 8.1 you must destroy or delete all copies of Software from all storage media in your possession.

9 Severability

In the event that any provision of this agreement is declared by any judicial or other competent authority to be void, voidable, illegal or otherwise unenforceable or indications of the same are received by either you or us from any relevant competent authority we shall amend that provision in such reasonable manner as achieves the intention of the parties without illegality, or at our discretion such provision may be severed from this agreement and the remaining provisions of this agreement shall remain in full force and effect.

10 Entire Agreement

You have read and understand this agreement and agree that it constitutes the complete and exclusive statement of the agreement between us with respect to the subject matter hereof and supersedes all proposals, representations, understandings and prior agreements, whether oral or written, and all other communications between us relating thereto.

11 Assignment

This agreement is personal to you (either an individual or a single corporate entity) and you may not assign, transfer, sub-contract or otherwise part with this agreement or any right or obligation under it without our prior written consent.

12 Law and Disputes

This agreement and all matters arising from it are governed by and construed in accordance with the laws of England and Wales, whose Courts shall have exclusive jurisdiction over all disputes arising in connection with this agreement.

If you have any questions about this agreement, write to us at Matrix Science Ltd., 64 Baker Street, London W1U 7GB, UK or call us at +44 (0)20 7486 1050 or email us at info@matrixscience.com

Xerces

/*
* The Apache Software License, Version 1.1
*
*
* Copyright (c) 1999-2001 The Apache Software Foundation. All rights
* reserved.
*
* Redistribution and use in source and binary forms, with or without
* modification, are permitted provided that the following conditions
* are met:
*
* 1. Redistributions of source code must retain the above copyright
* notice, this list of conditions and the following disclaimer.
*
* 2. Redistributions in binary form must reproduce the above copyright
* notice, this list of conditions and the following disclaimer in
* the documentation and/or other materials provided with the
* distribution.
*
* 3. The end-user documentation included with the redistribution,
* if any, must include the following acknowledgment:
* "This product includes software developed by the
* Apache Software Foundation (<http://www.apache.org/>)."
* Alternately, this acknowledgment may appear in the software itself,
* if and wherever such third-party acknowledgments normally appear.
*
* 4. The names "Xerces" and "Apache Software Foundation" must
* not be used to endorse or promote products derived from this
* software without prior written permission. For written
* permission, please contact apache@apache.org.
*
* 5. Products derived from this software may not be called "Apache",
* nor may "Apache" appear in their name, without prior written
* permission of the Apache Software Foundation.
*
* THIS SOFTWARE IS PROVIDED "AS IS" AND ANY EXPRESSED OR IMPLIED
* WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES
* OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE
* DISCLAIMED. IN NO EVENT SHALL THE APACHE SOFTWARE FOUNDATION OR
* ITS CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL,
* SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT
* LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF
* USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND
* ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY,
* OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT

* OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF
* SUCH DAMAGE.

* =====

*
*

* This software consists of voluntary contributions made by many
* individuals on behalf of the Apache Software Foundation and was
* originally based on software copyright (c) 1999, International
* Business Machines, Inc., <http://www.ibm.com>. For more
* information on the Apache Software Foundation, please see
* <http://www.apache.org/>.

*/

Curl

COPYRIGHT AND PERMISSION NOTICE

Copyright (c) 1996 - 2003, Daniel Stenberg, <daniel@haxx.se>.

All rights reserved.

Permission to use, copy, modify, and distribute this software for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Except as contained in this notice, the name of a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization of the copyright holder.

gzip, ht://Dig, cksum, touch, libstdc++

GNU GENERAL PUBLIC LICENSE Version 2, June 1991

Copyright (C) 1989, 1991 Free Software Foundation, Inc.
675 Mass Ave, Cambridge, MA 02139, USA
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software—to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation's software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you distribute copies of the software, or if you modify it.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

We protect your rights with two steps: (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.

Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we

want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

The precise terms and conditions for copying, distribution and modification follow.

GNU GENERAL PUBLIC LICENSE
TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The "Program", below, refers to any such program or work, and a "work based on the Program" means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term "modification".) Each licensee is addressed as "you".

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the Program is covered only if its contents constitute a work based on the Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.

1. You may copy and distribute verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and give any other recipients of the Program a copy of this License along with the Program.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

- a) You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.
- b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.
- c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.

In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

- a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

6. Each time you redistribute the Program (or any work based on the

Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

9. The Free Software Foundation may publish revised and/or new versions

of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and “any later version”, you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

NO WARRANTY

11. BECAUSE THE PROGRAM IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

12. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER

PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

bzip2

This program, "bzip2", the associated library "libbzip2", and all documentation, are copyright (C) 1996-2005 Julian R Seward. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.
3. Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.
4. The name of the author may not be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE AUTHOR "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Julian Seward, Cambridge, UK.

jseward@acm.org

bzip2/libbzip2 version 1.0.3 of 15 February 2005

SWIG

Simplified Wrapper and Interface Generator (SWIG)

SWIG is distributed under the following terms:

=====

I.

This software includes contributions that are Copyright (c) 1998-2002 University of Chicago. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution. Neither the name of the University of Chicago nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE UNIVERSITY OF CHICAGO AND CONTRIBUTORS

“AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE UNIVERSITY OF CHICAGO OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED

TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF

LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

II.

Copyright (c) 1995-1998
The University of Utah and the Regents of the University of California
All Rights Reserved

Permission is hereby granted, without written agreement and without license or royalty fees, to use, copy, modify, and distribute this software and its documentation for any purpose, provided that (1) The above copyright notice and the following two paragraphs appear in all copies of the source code and (2) redistributions including binaries reproduces these notices in the supporting documentation. Substantial modifications to this software may be copyrighted by their authors and need not follow the licensing terms described here, provided that the new terms are clearly indicated in all files where they apply.

IN NO EVENT SHALL THE AUTHOR, THE UNIVERSITY OF CALIFORNIA, THE UNIVERSITY OF UTAH OR DISTRIBUTORS OF THIS SOFTWARE BE LIABLE TO ANY PARTY FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OF THIS SOFTWARE AND ITS DOCUMENTATION, EVEN IF THE AUTHORS OR ANY OF THE ABOVE PARTIES HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

THE AUTHOR, THE UNIVERSITY OF CALIFORNIA, AND THE UNIVERSITY OF UTAH SPECIFICALLY DISCLAIM ANY WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE SOFTWARE PROVIDED HEREUNDER IS ON AN "AS IS" BASIS, AND THE AUTHORS AND DISTRIBUTORS HAVE NO OBLIGATION TO PROVIDE MAINTENANCE, SUPPORT, UPDATES, ENHANCEMENTS, OR MODIFICATIONS.

Linux glibc (section 6b applies)

Appendix G GNU Lesser General Public License

Version 2.1, February 1999

Copyright © 1991, 1999 Free Software Foundation, Inc.
59 Temple Place – Suite 330, Boston, MA 02111-1307, USA

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

[This is the first released version of the Lesser GPL. It also counts as the successor of the GNU Library Public License, version 2, hence the version number 2.1.]

G.0.1 Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public Licenses are intended to guarantee your freedom to share and change free software—to make sure the software is free for all its users.

This license, the Lesser General Public License, applies to some specially designated software—typically libraries—of the Free Software Foundation and other authors who decide to use it. You can use it too, but we suggest you first think carefully about whether this license or the ordinary General Public License is the better strategy to use in any particular case, based on the explanations below.

When we speak of free software, we are referring to freedom of use, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish); that you receive source code or can get it if you want it; that you can change the software and use pieces of it in new free programs; and that you are informed that you can do these things.

To protect your rights, we need to make restrictions that forbid distributors to deny you these rights or to ask you to surrender these rights. These restrictions translate to certain responsibilities for you if you distribute copies of the library or if you modify it.

For example, if you distribute copies of the library, whether gratis or for a fee, you must give the recipients all the rights that we gave you. You must make sure that they, too, receive or can get the source code. If you link other code with the library, you must provide complete object files to the recipients, so that they can relink them with the library after making changes to the library and recompiling it. And you must show them these terms so they know their rights.

We protect your rights with a two-step method: (1) we copyright the library, and (2) we offer you this license, which gives you legal permission to copy, distribute and/or modify the library.

To protect each distributor, we want to make it very clear that there is no warranty for the free library. Also, if the library is modified by someone else and passed on, the recipients should know that what they have is not the original version, so that the original author's reputation will not be affected by problems that might be introduced by others.

Finally, software patents pose a constant threat to the existence of any free program. We wish to make sure that a company cannot effectively restrict the users of a free program by obtaining a restrictive license from a patent holder. Therefore, we insist that any patent license obtained for a version of the library must be consistent with the full freedom of use specified in this license.

Most GNU software, including some libraries, is covered by the ordinary GNU General Public License. This license, the GNU Lesser General Public License, applies to certain designated libraries, and is quite different from the ordinary General Public License. We use this license for certain libraries in order to permit linking those libraries into non-free programs.

When a program is linked with a library, whether statically or using a shared library, the combination of the two is legally speaking a combined work, a derivative of the original library. The ordinary General Public License therefore permits such linking only if the entire combination fits its criteria of freedom. The Lesser General Public License permits more lax criteria for linking other code with the library.

We call this license the *Lesser* General Public License because it does *Less* to protect the user's freedom than the ordinary General Public License. It also provides other free software developers *Less* of an advantage over competing non-free programs. These disadvantages are

the reason we use the ordinary General Public License for many libraries. However, the Lesser license provides advantages in certain special circumstances.

For example, on rare occasions, there may be a special need to encourage the widest possible use of a certain library, so that it becomes a de-facto standard. To achieve this, non-free programs must be allowed to use the library. A more frequent case is that a free library does the same job as widely used non-free libraries. In this case, there is little to gain by limiting the free library to free software only, so we use the Lesser General Public License.

In other cases, permission to use a particular library in non-free programs enables a greater number of people to use a large body of free software. For example, permission to use the GNU C Library in non-free programs enables many more people to use the whole GNU operating system, as well as its variant, the GNU/Linux operating system.

Although the Lesser General Public License is Less protective of the users' freedom, it does ensure that the user of a program that is linked with the Library has the freedom and the wherewithal to run that program using a modified version of the Library.

The precise terms and conditions for copying, distribution and modification follow. Pay close attention to the difference between a "work based on the library" and a "work that uses the library". The former contains code derived from the library, whereas the latter must be combined with the library in order to run.

0. This License Agreement applies to any software library or other program which contains a notice placed by the copyright holder or other authorized party saying it may be distributed under the terms of this Lesser General Public License (also called "this License"). Each licensee is addressed as "you".

A "library" means a collection of software functions and/or data prepared so as to be conveniently linked with application programs (which use some of those functions and data) to form executables.

The "Library", below, refers to any such software library or work which has been distributed under these terms. A "work based on the Library" means either the Library or any derivative work under copyright law: that is to say, a work containing the Library or a portion of it, either verbatim or with modifications and/or translated straightforwardly into another language. (Hereinafter, translation is included without limitation in the term "modification".)

"Source code" for a work means the preferred form of the work for making modifications to it. For a library, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the library.

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running a program using the Library is not restricted, and output from such a program is covered only if its contents constitute a work based on the Library (independent of the use of the Library in a tool for writing it). Whether that is true depends on what the Library does and what the program that uses the Library does.

1. You may copy and distribute verbatim copies of the Library's complete source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and distribute a copy of this License along with the Library.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Library or any portion of it, thus forming a work based on the Library, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:
 - a. The modified work must itself be a software library.
 - b. You must cause the files modified to carry prominent notices stating that you changed the files and the date of any change.
 - c. You must cause the whole of the work to be licensed at no charge to all third parties under the terms of this License.
 - d. If a facility in the modified Library refers to a function or a table of data to be supplied by an application program that uses the facility, other than as an argument passed when the facility is invoked, then you must make a good faith effort to ensure that, in the event an application does not supply such function or table, the facility still operates, and performs whatever part of its purpose remains meaningful.

(For example, a function in a library to compute square roots has a purpose that is entirely well-defined independent of the application. Therefore, Subsection 2d requires that any application-supplied function or table used by this function must be optional: if the application does not supply it, the square root function must still compute square roots.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Library, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Library, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Library.

In addition, mere aggregation of another work not based on the Library with the Library (or with a work based on the Library) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may opt to apply the terms of the ordinary GNU General Public License instead of this License to a given copy of the Library. To do this, you must alter all the notices that refer to this License, so that they refer to the ordinary GNU General Public License, version 2, instead of to this License. (If a newer version than version 2 of the ordinary GNU General Public License has appeared, then you can specify that version instead if you wish.) Do not make any other change in these notices.

Once this change is made in a given copy, it is irreversible for that copy, so the ordinary GNU General Public License applies to all subsequent copies and derivative works made from that copy.

This option is useful when you wish to copy part of the code of the Library into a program that is not a library.

4. You may copy and distribute the Library (or a portion or derivative of it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange.

If distribution of object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place satisfies the requirement to distribute the source code, even though third parties are not compelled to copy the source along with the object code.

5. A program that contains no derivative of any portion of the Library, but is designed to work with the Library by being compiled or linked with it, is called a “work that uses the Library”. Such a work, in isolation, is not a derivative work of the Library, and therefore falls outside the scope of this License.

However, linking a “work that uses the Library” with the Library creates an executable that is a derivative of the Library (because it contains portions of the Library), rather than a “work that uses the library”. The executable is therefore covered by this License. Section 6 states terms for distribution of such executables.

When a “work that uses the Library” uses material from a header file that is part of the Library, the object code for the work may be a derivative work of the Library even though the source code is not. Whether this is true is especially significant if the work can be linked without the Library, or if the work is itself a library. The threshold for this to be true is not precisely defined by law.

If such an object file uses only numerical parameters, data structure layouts and accessors, and small macros and small inline functions (ten lines or less in length), then the use of the object file is unrestricted, regardless of whether it is legally a derivative work. (Executables containing this object code plus portions of the Library will still fall under Section 6.)

Otherwise, if the work is a derivative of the Library, you may distribute the object code for the work under the terms of Section 6. Any executables containing that work also fall under Section 6, whether or not they are linked directly with the Library itself.

6. As an exception to the Sections above, you may also combine or link a “work that uses the Library” with the Library to produce a work containing portions of the Library, and distribute that work under terms of your choice, provided that the terms permit modification of the work for the customer’s own use and reverse engineering for debugging such modifications.

You must give prominent notice with each copy of the work that the Library is used in it and that the Library and its use are covered by this License. You must supply a copy of this License. If the work during execution displays copyright notices, you must include the copyright notice for the Library among them, as well as a reference directing the user to the copy of this License. Also, you must do one of these things:

- a. Accompany the work with the complete corresponding machine-readable source code for the Library including whatever changes were used in the work (which must be distributed under Sections 1 and 2 above); and, if the work is an executable linked with the Library, with the complete machine-readable “work that uses the Library”, as object code and/or source code, so that the user can modify the Library and then relink to produce a modified executable containing the modified Library. (It is understood that the user who changes the contents of definitions files in the Library will not necessarily be able to recompile the application to use the modified definitions.)
- b. Use a suitable shared library mechanism for linking with the Library. A suitable mechanism is one that (1) uses at run time a copy of the library already present on the user’s computer system, rather than copying library functions into the executable, and (2) will operate properly with a modified version of the library, if the user installs one, as long as the modified version is interface-compatible with the version that the work was made with.
- c. Accompany the work with a written offer, valid for at least three years, to give the same user the materials specified in Subsection 6a, above, for a charge no more than the cost of performing this distribution.
- d. If distribution of the work is made by offering access to copy from a designated place, offer equivalent access to copy the above specified materials from the same place.
- e. Verify that the user has already received a copy of these materials or that you have already sent this user a copy.

For an executable, the required form of the “work that uses the Library” must include any data and utility programs needed for reproducing the executable from it. However, as a special exception, the materials to be distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

It may happen that this requirement contradicts the license restrictions of other proprietary libraries that do not normally accompany the operating system. Such a

contradiction means you cannot use both them and the Library together in an executable that you distribute.

7. You may place library facilities that are a work based on the Library side-by-side in a single library together with other library facilities not covered by this License, and distribute such a combined library, provided that the separate distribution of the work based on the Library and of the other library facilities is otherwise permitted, and provided that you do these two things:
 - a. Accompany the combined library with a copy of the same work based on the Library, uncombined with any other library facilities. This must be distributed under the terms of the Sections above.
 - b. Give prominent notice with the combined library of the fact that part of it is a work based on the Library, and explaining where to find the accompanying uncombined form of the same work.
8. You may not copy, modify, sublicense, link with, or distribute the Library except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, link with, or distribute the Library is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.
9. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Library or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Library (or any work based on the Library), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Library or works based on it.
10. Each time you redistribute the Library (or any work based on the Library), the recipient automatically receives a license from the original licensor to copy, distribute, link with or modify the Library subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties with this License.
11. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Library at all. For example, if a patent license would not permit royalty-free redistribution of the Library by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Library.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply, and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

12. If the distribution and/or use of the Library is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Library under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

13. The Free Software Foundation may publish revised and/or new versions of the Lesser General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Library specifies a version number of this License which applies to it and “any later version”, you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Library does not specify a license version number, you may choose any version ever published by the Free Software Foundation.

14. If you wish to incorporate parts of the Library into other free programs whose distribution conditions are incompatible with these, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

15. BECAUSE THE LIBRARY IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE LIBRARY, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE LIBRARY “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE LIBRARY IS WITH YOU. SHOULD THE LIBRARY

PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE LIBRARY AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE LIBRARY (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE LIBRARY TO OPERATE WITH ANY OTHER SOFTWARE), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

G.0.2 How to Apply These Terms to Your New Libraries

If you develop a new library, and you want it to be of the greatest possible use to the public, we recommend making it free software that everyone can redistribute and change. You can do so by permitting redistribution under these terms (or, alternatively, under the terms of the ordinary General Public License).

To apply these terms, attach the following notices to the library. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the “copyright” line and a pointer to where the full notice is found.

one line to give the library's name and an idea of what it does.
Copyright (C) year name of author

This library is free software; you can redistribute it and/or modify it under the terms of the GNU Lesser General Public License as published by the Free Software Foundation; either version 2.1 of the License, or (at your option) any later version.

This library is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU Lesser General Public License for more details.

You should have received a copy of the GNU Lesser General Public License along with this library; if not, write to the Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02111-1307, USA.

Also add information on how to contact you by electronic and paper mail.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a “copyright disclaimer” for the library, if necessary. Here is a sample; alter the names:

Yoyodyne, Inc., hereby disclaims all copyright interest in the library ‘Frob’ (a library for tweaking knobs) written by James Random Hacker.

signature of Ty Coon, 1 April 1990
Ty Coon, President of Vice

That's all there is to it!

Regex

Copyright 1992, 1993, 1994 Henry Spencer. All rights reserved. This software is not subject to any license of the American Telephone and Telegraph Company or of the Regents of the University of California.

Permission is granted to anyone to use this software for any purpose on any computer system, and to alter it and redistribute it, subject to the following restrictions:

1. The author is not responsible for the consequences of use of this software, no matter how awful, even if they arise from flaws in it.
2. The origin of this software must not be misrepresented, either by explicit claim or by omission. Since few users ever read sources, credits must appear in the documentation.
3. Altered versions must be plainly marked as such, and must not be misrepresented as being the original software. Since few users ever read sources, credits must appear in the documentation.
4. This notice may not be removed or altered.

```
=====  
== /*-  
* Copyright (c) 1994  
* The Regents of the University of California. All rights reserved.  
*  
* Redistribution and use in source and binary forms, with or without  
* modification, are permitted provided that the following conditions  
* are met:  
* 1. Redistributions of source code must retain the above copyright  
* notice, this list of conditions and the following disclaimer.  
* 2. Redistributions in binary form must reproduce the above copy-  
right  
* notice, this list of conditions and the following disclaimer in  
the  
* documentation and/or other materials provided with the distribu-  
tion.
```

* 3. All advertising materials mentioning features or use of this software
* must display the following acknowledgement:
* This product includes software developed by the University of
* California, Berkeley and its contributors.
* 4. Neither the name of the University nor the names of its contributors
* may be used to endorse or promote products derived from this software
* without specific prior written permission.
*
* THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS "AS IS"
AND
* ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO,
THE
* IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR
PURPOSE
* ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE
LIABLE
* FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL
DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE
GOODS
* OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION)
* HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT,
STRICT
* LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN
ANY WAY
* OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY
OF
* SUCH DAMAGE.
*
* @(#)COPYRIGHT 8.1 (Berkeley) 3/16/94
*/

PLPA

Copyright (c) 2004-2006 The Trustees of Indiana University and Indiana
University Research and Technology
Corporation. All rights reserved.
Copyright (c) 2004-2005 The Regents of the University of California.
All rights reserved.
Copyright (c) 2007 Cisco Systems, Inc. All rights reserved.
Portions copyright:

Copyright (c) 2004-2005 The University of Tennessee and The University

of Tennessee Research Foundation. All rights reserved.

Copyright (c) 2004-2005 High Performance Computing Center Stuttgart, University of Stuttgart. All rights reserved.

Copyright (c) 2006, 2007 Advanced Micro Devices, Inc. All rights reserved.

\$COPYRIGHT\$

Additional copyrights may follow

\$HEADER\$

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer listed in this license in the documentation and/or other materials provided with the distribution.
- Neither the name of the copyright holders nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

The copyright holders provide no reassurances that the source code provided does not infringe any patent, copyright, or any other intellectual property rights of third parties. The copyright holders disclaim any liability to any recipient for claims brought against recipient by any third party for infringement of that parties intellectual property rights.

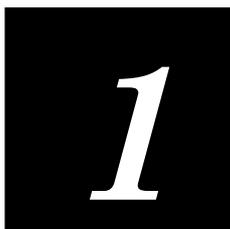
THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Contents

<i>Introduction</i>	1
<i>Installation: Unix</i>	5
<i>Installation: Microsoft Windows</i>	25
<i>Validation</i>	49
<i>Sequence Database Setup</i>	53
<i>Configuration & Log Files</i>	83
<i>Program Reference</i>	123
<i>I/O File Formats</i>	147
<i>Taxonomy</i>	167
<i>Mascot Daemon</i>	183
<i>Cluster Mode</i>	185
<i>Security</i>	217
<i>Basic Regular Expressions</i>	227
<i>Error Messages</i>	231
<i>System Limits</i>	233
<i>Web Server Configuration</i>	235
<i>Web Browsers</i>	245

Typographical Conventions

Description	Example
<p>Filenames, pathnames, directory names (folders), programs, and commands are printed in italic fixed pitch font. In Unix, these names are case sensitive.</p>	<pre><i>nph-mascot.exe</i> <i>/usr/local/httpd</i></pre>
<p>The contents of text files are printed in fixed pitch font on a grey background. Where it is necessary to break a long line, this is indicated by an indent and the symbols ↵ ↵. Text omitted from a line is indicated by an ellipsis ..., while missing lines are indicated by 3 vertical periods.</p>	<pre>>owl 100K_RAT 100 KD PROTEIN ↵ ↵(EC 6.3.2.-). >owl 100K_RAT ... NORVEGICUS (RAT).</pre>
<p>Text which should be entered literally is shown in bold fixed pitch font on a grey background. Control characters are shown in angle brackets, apart from <return> (i.e. carriage return, newline, or enter) which is only shown in ambiguous cases. Bold italic fixed pitch font indicates a variable for which an appropriate value should be entered.</p>	<pre>C:\TEMP> ftp ncbi.nlm.nih.gov % <i>kill PID</i></pre>



Introduction

Mascot is a software system for protein identification by matching mass spectrometry (MS) data against FASTA format protein or nucleic acid sequence databases. This can be done in three different ways:

1. A Peptide Mass Fingerprint (PMF), in which the MS data are peptide molecular weights from the digestion of a protein by an enzyme.
2. A Sequence Query (SQ), also called a sequence tag, in which MS data are combined with amino acid sequence or composition data.
3. An MS/MS Ions Search (MIS), which uses MS/MS data from one or more peptides.

MS data are submitted to Mascot in the form of peak lists. That is, lists of centroided mass values, possibly with associated intensity values. Intensity information is not required for a peptide mass fingerprint, but is very important in an MS/MS ions search. The result of a search is a ranked list of the most closely matching proteins. Mascot uses a probability based scoring algorithm, so that it is possible to report whether a match is statistically significant. If an exact match is not present in the database, the highest scoring matches will be those entries which exhibit the greatest homology.

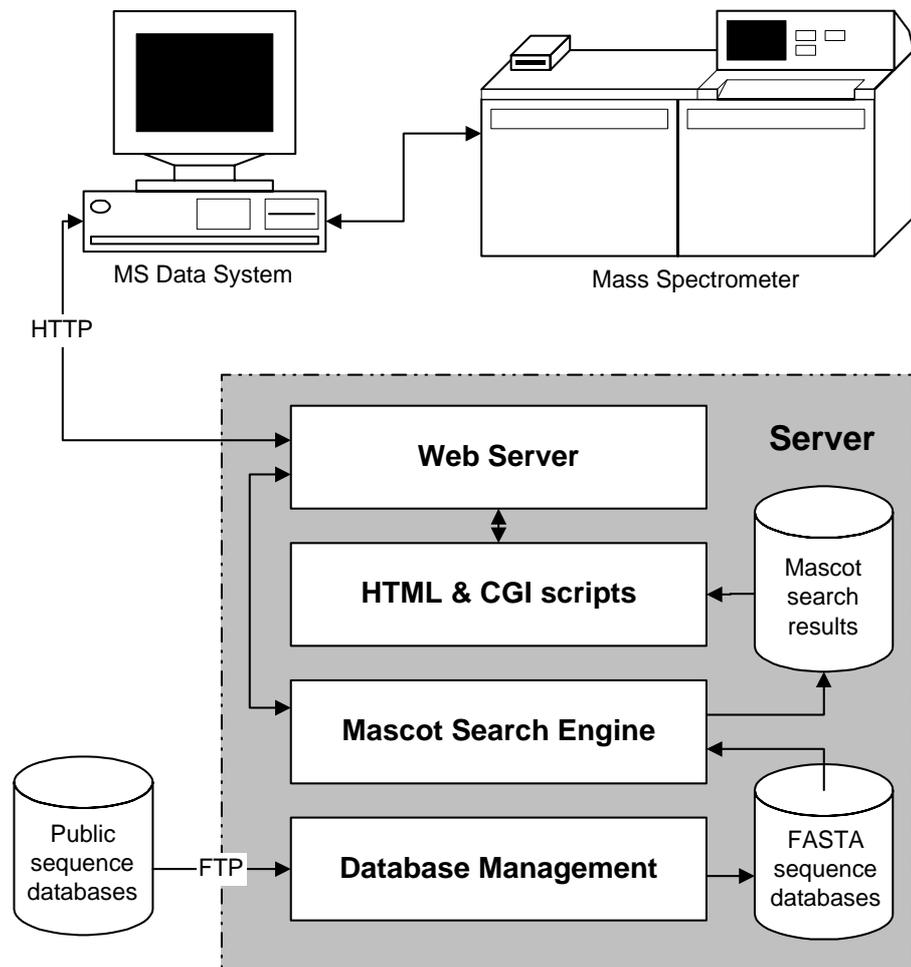
Overview

This manual describes how to install, configure and administer Mascot. It is not a User Guide. Mascot includes a linked collection of HTML help pages that provide guidance and application related reference material for end-users.

Mascot conforms to a client / server architecture, and the primary user interface is a web browser. Any JavaScript aware browser is suitable, though a version 4 or later browser is recommended. Searches are defined using HTML forms. A form may be used to input search parameters and data and may also specify a local text file to be uploaded to the server. This uploaded file will normally contain data, but may also contain additional search parameters.

2 Mascot: Installation and Setup

The Mascot search engine is executed as a CGI program. On completion of a search, it calls a Perl CGI script that reads the results file and returns an HTML report to the client browser. Links to additional CGI scripts provide more detailed views of the results.



Mascot Components

In this manual, “server” refers to the data system on which the Mascot search engine executes. The term “client” is used very loosely. It may refer to a data system attached to a mass spectrometer, or it may refer to

any system at which a user interacts with the Mascot server via a web browser.

In a small laboratory, the server and client may be one and the same computer. This doesn't affect installing or using Mascot, but it does introduce additional considerations, such as the need to adjust system priorities to ensure that the instrument control and data acquisition software is responsive to the real-time needs of both instrument and operator.

Configuration

Mascot configuration files are structured text files. Modifications can be made using a browser-based configuration editor and take effect without a system restart.

Search Engine

The Mascot search engine accepts data and parameters on STDIN in MIME format, executes a search of the specified FASTA format database, and outputs a structured text file containing the search results together with the input data and the complete set of search parameters.

The results file contains everything necessary to repeat the search at a later date, should the need arise. In the default configuration, a new directory is created on the server for each day's results files. If required, the contents of these results files can be parsed into an external database to be queried and analysed.

Monitor

Swapping databases without disrupting ongoing searches is handled by Mascot Monitor. The new database is compressed and tested by running a standard search. If errors are detected in the new database, the database exchange process is abandoned, and searches continue to use the earlier database

Assuming the test is successful, all new searches are performed against the new database, while searches that were in progress against the old database are allowed to continue. Once the final search against the old database is complete, the compressed files are deleted and the FASTA file is moved to an archive directory. If the database being exchanged is memory mapped, the mapping and unmapping are also handled automatically.

Status

The Mascot package includes a CGI application that provides a live status display via a web browser. For each database, the Mascot job queue, the executing jobs, and the completed jobs are listed. The status lines for completed jobs contain hyperlinks to individual results reports.

Review

Review is a CGI application that provides easy access to the flat file database of search result files. Key search parameters, such as time and date, job number, user name, search type, etc. are displayed in a spreadsheet-like table. Columns can be hidden, sorted and filtered to facilitate locating a specific file or group of files. Each row includes hyperlinks, either to generate a Mascot results reports or to display the file contents as raw text.

2

Installation: Unix

Release Notes

Refer to the release notes for last-minute additions and platform specific notes.

Cluster Mode

If you have a licence to run Mascot on four or more processors, and plan to do so on a networked cluster of machines, then please familiarise yourself with the material in Chapter 11, Cluster Mode, before proceeding with the installation.

System Requirements

Web Server

Mascot is compatible with most web servers, and we do not recommend any specific product. Appendix D provides configuration information for some of the more widely used servers.

If a web server is being installed for the first time, in connection with the installation of Mascot, it is essential to verify that everything is working correctly before attempting to install Mascot.

Perl

Mascot requires Perl, together with several Perl library modules, some of which may not be included in a standard Perl distribution. Perl 5.8 is recommended. Earlier versions are supported on some platforms; refer to the release notes for details.

The latest versions of Perl and Perl library modules can be downloaded from any site in the CPAN network:

<http://www.cpan.org/>

Although binary distributions are available for many platforms and operating systems, it may be preferable to download the source distribution for compilation on the target system.

Mascot scripts assume that Perl can be found at `/usr/local/bin/perl`. If Perl is installed in a different path, just add a soft link:

```
ln -s /actual/location/of/perl /usr/local/bin/perl
```

If any library modules are missing, this will be identified during the installation procedure. For convenience, source distribution packages for required library modules are included on the Mascot CD-ROM.

All the Perl modules conform to a standard compilation and installation procedure:

```
perl Makefile.PL  
make  
make test  
make install
```

On some platforms, the make utility is called `gmake`.

HTTP and LWP

If the installation script reports that either `HTTP::Request::Common` or `LWP::UserAgent` is missing, the following modules should be installed in the order listed:

```
MIME.TAR (MIME-Base64-2.11.tar)  
URI.TAR (URI-1.06.tar)  
HTML.TAR (HTML-Parser-3.07.tar)  
LIBNET.TAR (libnet-1.0703.tar)  
MD5.TAR (Digest-MD5-2.09.tar)  
LWP.TAR (libwww-perl-5.48.tar)
```

GD

The GD module is a Perl interface to Thomas Boutell's libgd library. It can be downloaded from any CPAN site:

```
http://search.cpan.org/dist/GD/
```

The current version of GD.pm (2.35) requires Perl version 5.6 or higher and the following libraries:

The libgd graphics library:

```
http://www.boutell.com/gd/
```

The PNG graphics library:

<http://www.libpng.org/pub/png/libpng.html>

The zlib compression library:

<http://www.gzip.org/zlib/>

Early versions of GD and libgd supported GIF format graphics. In later versions, this support was removed for licensing reasons and replaced with support for PNG format graphics. Mascot scripts automatically switch between GIF and PNG formats, so you are free to use earlier versions of GD if other applications require GIF support.

Compiling, linking and installing the support libraries for PNG can be difficult and time consuming. For this reason, version 1.19 of GD, which supports the GIF format and does not require external libraries, is supplied on the Mascot CD-ROM or can be downloaded from CPAN.

GD.TAR (GD-1.19.tar)

Mascot Directory Structure

There are two directory structures to consider. One consists of the “real” paths to files on disk, the other consists of the “virtual” directories which define the web server URL’s. The virtual directories are mapped to real directories. For example, the server URL

<http://your.domain/mascot/index.html>

might be mapped to the disk file

/usr/local/mascot/html/index.html

Any virtual directory that contains CGI executable programs (e.g. *nph-mascot.exe*) or scripts (e.g. *master_results.pl*) must have script execution enabled.

Under normal circumstances, if a directory is mapped to a URL, all of its subdirectories are also accessible as subdirectories of the URL. Figure 2.1 shows the recommended directory structure for Mascot. The root of this structure can be any convenient path.

Some of the directory paths can be changed by modifying the configuration file, *mascot.dat*. For example, it may be desirable to have the FASTA database files on a separate disk from the rest of the files. Care should be taken with any changes which affect a URL mapped directory or file, because this may require one or more HTML files to be edited to modify links.

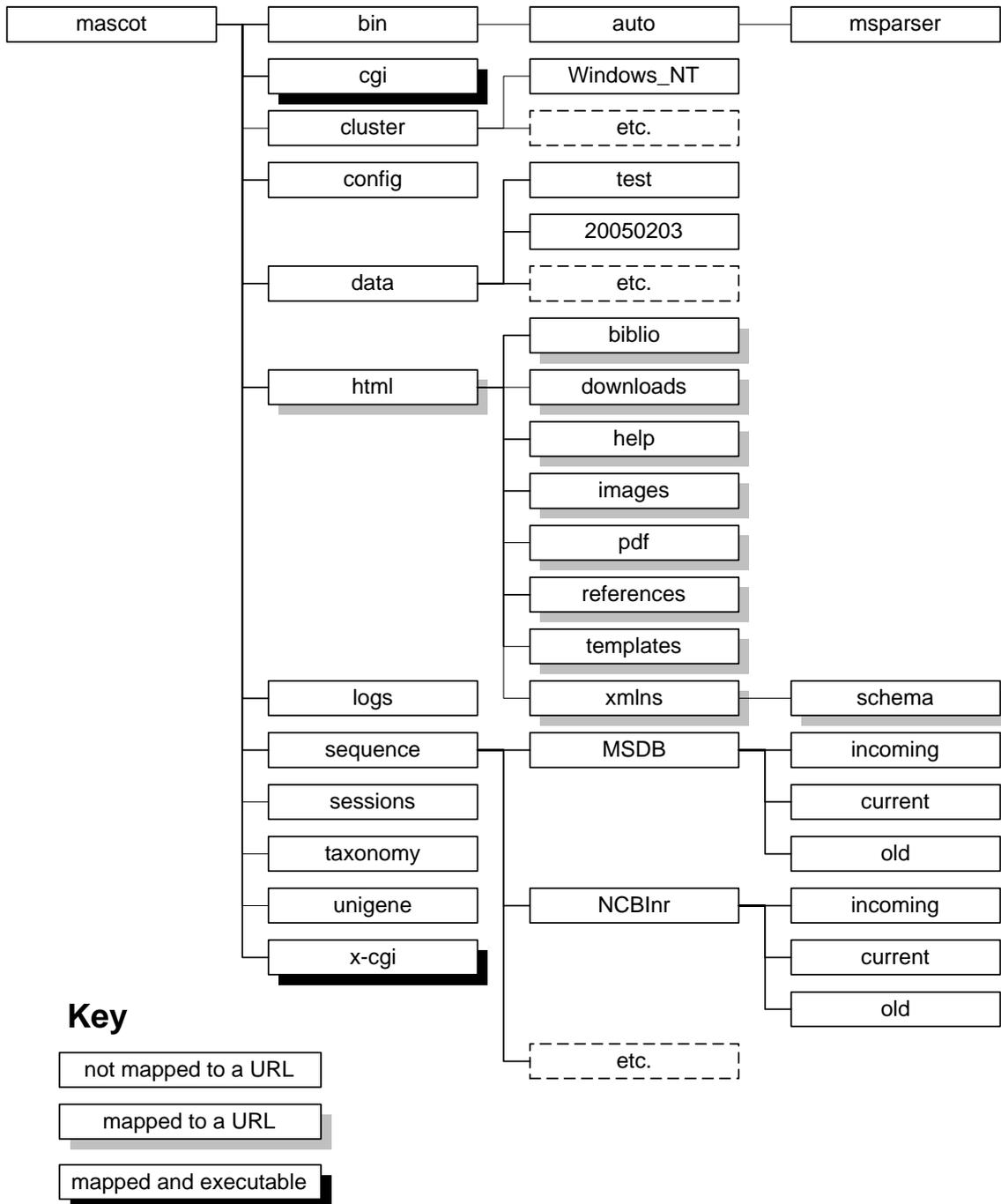


Figure 2.1 Mascot Directory Structure

In most cases, the contents of the directories can be deduced from their names:

bin contains (non-CGI) executables.

cgi contains CGI executables

cluster contains one or more sub-directories for platform specific executables, for distribution to the nodes in a cluster

config contains configuration files

data contains Mascot results files. By default, a new sub-directory is created for each day's results files. The name of each sub-directory is that day's date in ISO format, *yyyymmdd*.

html is the root directory for documents

logs contains search and error logs, etc.

sequence contains a sub-directory for each FASTA database. As illustrated, for each database there are 3 sub-directories to organise the FASTA files into new downloads (*incoming*), active databases (*current*) and the most recently replaced files (*old*).

sessions contains security session files

taxonomy contains taxonomy resources

unigene contains sub-directories for species specific UniGene indexes

x-cgi is a directory for administrative CGI executables, to which access may need to be restricted. This can be achieved using either Mascot security or web server security.

Installation

SGI Irix installations only: Mascot must be installed on a local drive. This is because it is not possible to *chcap* files on NFS mounts. This limitation only applies to the program files. Result files and sequence database files can reside on remote drives.

Unpack the Mascot file system

Create a new directory called *mascot*. This directory should *not* be in a path mapped to a web server URL.

Copy the files *mascot.tar.bz2* and *swissprot.tar.bz2* from the Mascot CD-ROM into the new *mascot* directory, and unpack the archives:

```
bzip2 -d mascot.tar.bz2
bzip2 -d swissprot.tar.bz2
tar xvf mascot.tar
tar xvf swissprot.tar
```

For 64 bit Linux, the 64 bit binaries should be unpacked after mascot.tar:

```
bzip2 -d mascot-64.tar.bz2
tar xvf mascot-64.tar
```

To conserve disk space, you can pipe the output from decompression direct to tar. For example:

```
bzip2 -dc mascot.tar.bz2 | tar xvf -
```

This will create the directory structure illustrated in Figure 2.1.

Certain Mascot files and directories must be writeable by the user who owns the web server process. Within the distribution archive, the ownership is 60001:60001, which usually corresponds to nobody:nobody. If your web server runs under a different UID, then use the `chown` command to change the ownership of the Mascot files and directories. For example, the ownership under Linux may be `apache:apache`:

```
chown -R apache:apache /usr/local/mascot/*
```

(If this is not acceptable, then the `logs`, `config`, `sessions`, and `data` directories, plus the file `logs/errorlog.txt` must be made writeable by the web server process).

SGI Irix installations only: If you are installing Mascot for SGI Irix, all the binary executables (files with `.exe` extensions) must have their capabilities changed using the `chcap` command. While logged in as root, you must enter the following command at a shell prompt in the `mascot/bin`, `mascot/cgi` and `mascot/x-cgi` directories:

```
chcap CAP_SCHED_MGT+ep *.exe
```

Licence file

Operation of Mascot is only possible with a valid licence, normally supplied by email. Before proceeding with the installation, you must copy your licence file into the `mascot/config` directory and (if necessary) rename it to `mascot.license`.

Create URL mappings

Add the following mappings to your web server configuration, (substituting your actual disk path to the new `mascot` directory):

Disk path	URL	Executable
<code>/usr/local/mascot/cgi</code>	<code>/mascot/cgi</code>	Yes
<code>/usr/local/mascot/html</code>	<code>/mascot</code>	No
<code>/usr/local/mascot/x-cgi</code>	<code>/mascot/x-cgi</code>	Yes

You may wish to restrict access to the administrative programs by setting a password or IP address restriction on `/mascot/x-cgi`.

Different web servers support different methods of adding URL mappings. Commercial servers often have a graphical user interface for administration tasks. Some notes on web server configuration are provided later in Appendix D.

Installation Script

Step 1: Web Server Operation

Launch a JavaScript aware web browser, and navigate to the URL corresponding to `install.html`:

`http://your.domain/mascot/install.html`

Follow the instructions on this web page and those that follow to perform some simple system checks and customise the Mascot configuration file (`mascot.dat`).

Step 2: Perl

If you get an error message, or a “File Save As...” dialog box, after clicking on the “Test Perl” button, then Perl is not functioning correctly. This must be corrected before proceeding. Possible reasons for this problem are listed along with useful links:

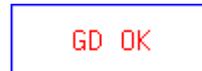
- Perl is not installed, or was installed incorrectly.
- Perl, or a soft link to Perl, was not found at `/usr/local/bin/perl`
- This is an early version of Perl, which does not include the CGI or File modules
- The `mascot/cgi` directory is not configured for CGI execution.

- You are not using a JavaScript aware web browser.

If Perl is functioning correctly, the next page displays the Perl version number. If any of the required modules are missing, there will be a warning. Instructions for installing the required modules can be found towards the beginning of this chapter.

Step 3: GD Graphics Library

Assuming Perl and the required modules are present, click on the button to test the GD Graphics library. If GD is installed and working, the next page contains a small graphic to confirm this:



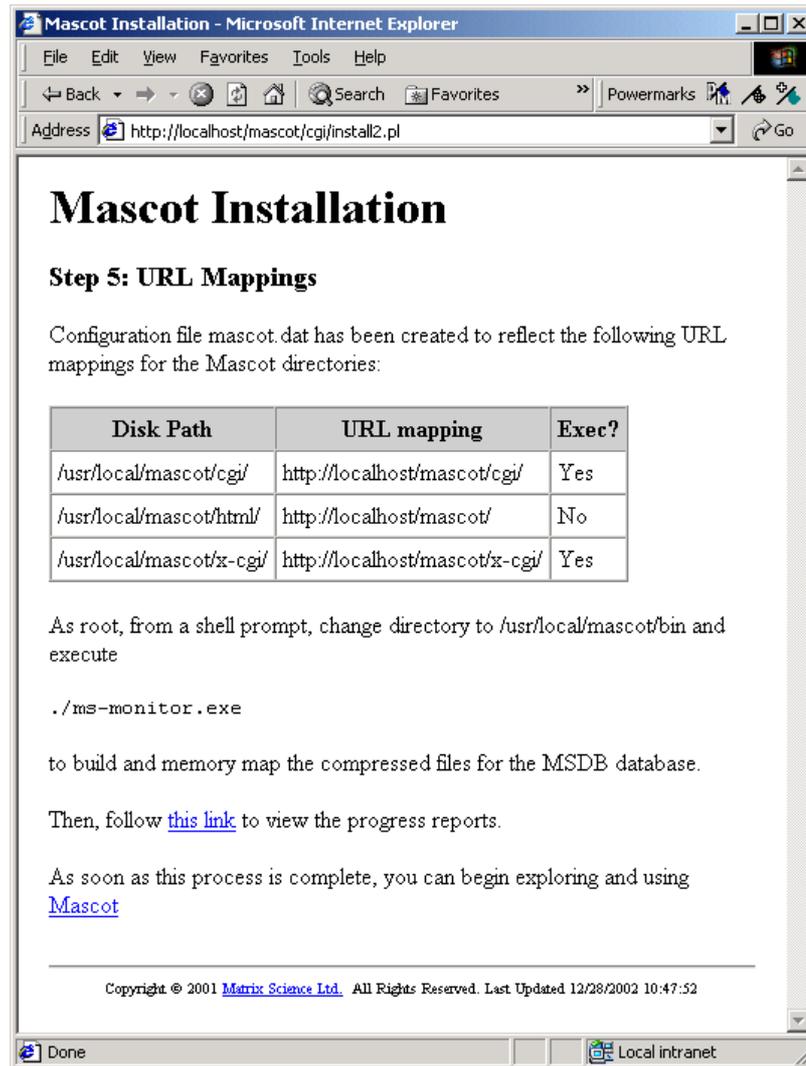
If you do not see this picture, or get an error message, refer to the first part of this chapter for information on downloading GD.pm.

Step 4: Configuration



Indicate whether you plan to configure Mascot as a single (SMP) server or a cluster. Enter the full path to the mascot directory and press 'Configure Mascot'

Step 5: URL Mappings



If you chose to configure Mascot as a single (SMP) server, you will see a screen similar to the one above, and can proceed to start Mascot Monitor. If you chose cluster mode, refer to Chapter 11 for further configuration information.

Start Mascot Monitor

A copy of the SwissProt database is included in the files copied from the CD-ROM. It is recommended that the operation of Mascot is verified and tested using this database before making any modifications to the directory structure or configuration files

Mascot Monitor (*ms-monitor.exe*) is used to manage the swapping and memory mapping of the sequence databases used by Mascot. For Mascot to operate, *ms-monitor.exe* must be running at all times.

In order to test the Mascot installation, start Monitor at a shell prompt. (You must have root privileges for this)

```
cd /usr/local/mascot/bin
./ms-monitor.exe
```

The output from Mascot Monitor is also appended to the monitor.log file.

Normally, *ms-monitor.exe* should be started as part of the normal server boot process. The method of doing this is operating system dependent. In the case of Irix, for example, */etc/rc2.d/S90monitor* could be a link to a very simple shell script called */etc/init.d/monitor*:

```
#!/bin/sh

MASCOT_HOME = /usr/local/mascot

if [ -x $MASCOT_HOME/bin/ms-monitor.exe ]; then
  cd $MASCOT_HOME/bin
  ./ms-monitor.exe > /dev/null
  echo "starting Mascot monitor"
fi
```

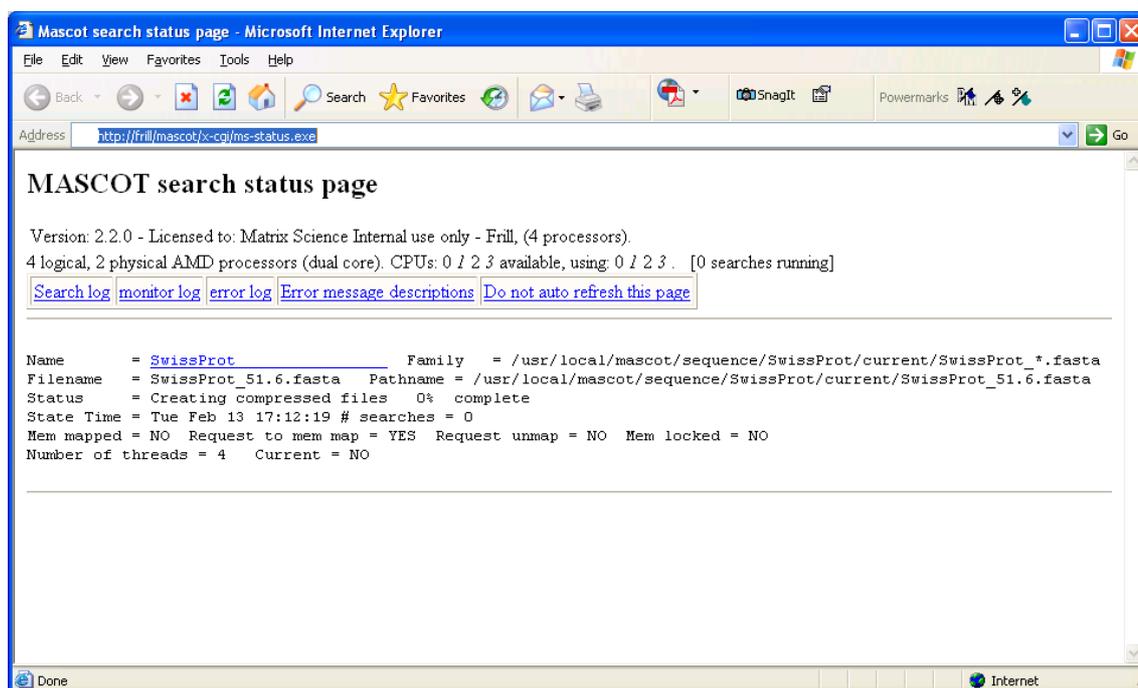
SGI Irix only: It is possible to run *ms-monitor.exe* from a user ID that is not root. To do this, you must change the 'capabilities' of *ms-monitor.exe* while logged on as root:

```
chcap CAP_MEMORY_MGT+ep ms-monitor.exe
chcap CAP_MEMORY_MGT+ep ms-lockmem.exe
```

It will then be possible for any user with the appropriate directory and file access rights to run *ms-monitor.exe*.

Verify System Operation

Once this Monitor has been started, use the link on the final web page to invoke Mascot Status. You should see a display resembling the following:



If an error occurs, use the links to the monitor log and the error log to investigate the cause. If all is well, you will see the following messages displayed on the status line for SwissProt:

```

Creating compressed files
Running 1st test
First test just run OK
Trying to memory map files
Just enabled memory mapping
In Use

```

You can begin exploring and using Mascot. However, do not try to run searches or view results reports until the relevant sequence database is 'In Use'.

Security

Mascot security is disabled on installation. To enable Mascot security, refer to Chapter 12

Keyword Indexing

Users of Mascot may wish to be able to search the help text by keywords or phrases. A suitable indexing tool, `ht://Dig`, is included on the Mascot CD-ROM. If you choose to use a different search engine, the search field at the top of each Mascot HTML page will require modification.

A C++ compiler is required to compile the `ht://Dig` search engine. If you do not have a suitable compiler, you may be able to download binaries from <http://www.htdig.org/files/binaries/>

Full instructions for compiling and installing `ht://Dig` can be found in the distribution archive. Alternatively, the following instructions can be followed for a quick, standard installation:

Copy the archive `htdig-3.1.6.tar` to a temporary location and unpack it

```
tar -xvf htdig-3.1.6.tar
```

Change to the new `htdig-3.1.6` directory and create a configuration file for a suitable installation path. For example:

```
cd htdig-3.1.6
```

```
./configure --prefix=/usr/local/htdig
```

Edit the new `CONFIG` file to reflect your Mascot directory structure and URL mappings:

```
CGIBIN_DIR = /usr/local/mascot/cgi
```

```
IMAGE_DIR = /usr/local/mascot/html/images
```

```
IMAGE_URL_PREFIX = /mascot/images
```

```
SEARCH_DIR = /usr/local/mascot/html
```

Execute the make files

```
make depend
```

```
make
```

```
make install
```

Edit the following values in the `ht://Dig` configuration file, `/usr/local/htdig/conf/htdig.conf`

```
start_url: http://your_host/mascot/home.html
limit_urls_to: /mascot/
exclude_urls: .pl .exe .gif .jpg .pdf
```

Finally, build an index of the Mascot web site documents:

```
/usr/local/htdig/bin/rundig -v
```

Indexing will only take a minute or two. Use of the `-v` flag causes verbose progress reports to be generated.

Manual Installation

If you prefer not to use the installation scripts to customise *mascot.dat*, open *mascot/config/not.mascot.dat* in a text editor and make the following changes:

1. Change every occurrence of

```
###ROOT###
```

to the path to the Mascot directory, e.g.

```
/usr/local/mascot
```

2. Change the value of threads for each database to the number of licensed cores.

NB If this is the master system in a cluster, then every occurrence of threads should be set to 1.

3. Change every occurrence of

```
###URL###
```

to the URL of the Mascot server, up to and including the mascot directory, e.g.

```
http://your_host/mascot
```

4. Change every occurrence of

```
###HOSTNAME###
```

to the host name of the Mascot server, optionally qualified with the domain name, e.g.

```
your_host or your_host.your_domain
```

5. If your web server is Apache, change the value of the `ForkForUnixApache` parameter from 0 to 1.

6. Save the modified file as *mascot.dat* and continue the installation procedure at 'Start Mascot Monitor'.

7. Change to the directory *mascot/bin/auto/msparser*. You will find multiple **.so* files. Identify the correct file for your version of Perl. The file names contain the code '50' for Perl 5.005, '56' for Perl 5.6, and '58' for Perl 5.8. For Perl compiled with support for threads, the file names contain the code 'thread'. For Perl compiled with support for multiplicity, the file names contain the code 'multi'. For 64 bit Perl, the file names contain the code '_64'. Having identified the correct file for your system, create a link called *msparser.so*. For example

```
ln -s msparser58-thread-multi.so msparser.so
```

The latest version of *msparser.pm* will have been unpacked into the *mascot/bin* directory. If you already have an earlier version of Mascot Parser installed on another path, you should update this to avoid potential confusion.

Hyper-threading

Linux on Intel only: Hyper-threading is a technique used by Intel to improve the performance of multi-threaded programs by putting multiple 'processors' onto the same physical chip. Pentium 4 Xeon processors include this functionality. The performance does not increase linearly with the number processors, (currently limited to 2), because the individual processor cores share other resources such as the on-chip cache. In many cases, there is no performance improvement for Mascot searches when hyper-threading is enabled. On most systems, a BIOS setting can be used to enable and disable hyper-threading.

Hyper-threading is detected automatically and the additional processors may be used to increase performance. For example, with a 2 CPU Mascot license on a dual processor Xeon system, there will appear to be 4 processors. On this system, the Mascot search status screen will display:

```
4 processors available, [Hyper-threading enabled], (numbers 0 1 2 3 ).
Using processor numbers 0 1 2 3.
```

(Processor numbers 2 and 3 are in italics to show that they are the 'hyper-threaded' cpus)

On the same system, with a single processor license, the Mascot search status screen will display:

```
4 processors available, [Hyper-threading enabled], (numbers 0 1 2 3 ).
Using processor numbers 0 2.
```

Processor numbers 0 and 2 are chosen by default because they are on the same physical chip. Attempting to set `Processors=0,1` in the `processors` section of `mascot.dat` will fail.

To make use of these other processors, use the database maintenance utility to change the number of threads to the number of logical processors. For example, on a dual CPU Xeon system, change the number from 2 to 4.

Under Linux, Mascot can only detect hyper-threading when using kernel version 2.4.17 and later. For earlier versions of the kernel, hyper-threading must be disabled in the BIOS for Mascot to run. A 2.6 kernel is usually required if the number of physical processors (sockets) on the system is greater than the number of Mascot licenses, refer to the release notes for further details.

Troubleshooting

System limits

Memory limits

There are several types of memory limits that can stop Mascot from running:

1. Virtual address space. When files are memory mapped, the address space required can be large – the amount of physical RAM / swap space is not an issue here.
2. The amount of memory that can be locked. On most systems, memory can only be locked by root.
3. Physical memory. It is obviously not possible to lock more memory than is physically available!
4. Data space. The amount of memory that an executable or Perl script has access to. The default is sometimes too small to run `master_results.pl`, and big searches.
5. Swap space. May need to be set to at least 2 Gb for very large searches.
6. Stack space. Not normally an issue for executables or any of the perl scripts.
7. Thread stack space. Not normally an issue for executables. The perl scripts are not threaded

File size limits.

This is normally unlimited, but note that under IBM AIX, file size is limited to 1Gb by default. To change this limit, edit `/etc/security/limits` and set `fsize` to `-1`.

You should manually verify that your system can successfully FTP a file larger than 2 Gb, as FTP doesn't necessarily report an error when it fails.

How the errors are reported

If the Mascot executables report a memory error, the error can be found in the `errorlog.txt` file, including the error code returned by the operating system. For a Perl script running in CGI mode, the web server may just kill the job, and no error will be logged.

Determining what the limits are.

Most systems have two sets of limits – the current limits and the hard limits.

There is no standard Unix command across all platforms, although 'limit' or 'limits' will work on most systems from the 'C' shell – if the command doesn't work, try typing 'csh' first. (Also, there is generally no man page for limit – the information can be found in the csh man page – this is because it obviously has to be a built in command for the shell)

```
$ limit
cputime      unlimited
filesize    unlimited
datasize    1048576 kbytes
stacksize   65536 kbytes
coredumpsize unlimited
memoryuse   250004 kbytes
descriptors 200
vmemory     1048576 kbytes
threads     1024
```

```
$ limit -h
cputime      unlimited
filesize    unlimited
datasize    1048576 kbytes
stacksize   524288 kbytes
coredumpsize unlimited
memoryuse   524288 kbytes
descriptors 2500
vmemory     1048576 kbytes
threads     1024
```

These values will be different for root and a normal user, and possibly different again for 'nobody'. Since you mostly cannot log in as nobody, it can be hard to find out what the real values are. If a script or binary is failing in the web browser, try running from the command line as both root and a normal user.

Changing the default limits

There are different utilities / configuration files on every system. Refer to system documentation.

Detailed Information on each memory limit

This section gives details about how the mascot software reports errors, and tries to increase the limits where appropriate.

Virtual address space

On certain platforms, the executables are compiled as both 32bit and 64 bit programs. If you use the 32 bit executables, the amount of virtual address space is limited to 2,147,483,648 bytes, (2Gb), and this limit cannot be exceeded. However, default limit are often set lower than this.

Just before a file is memory mapped, Mascot Monitor sets the default limit to the hard limit. Under IRIX this is done using

```
getrlimit(RLIMIT_VMEM, &rlp)
rlp.rlim_cur = rlp.rlim_max;
setrlimit(RLIMIT_VMEM, &rlp);
```

To restrict the amount of address space claimed, add an entry

```
Vmemory xxxx
```

to the options section of mascot.dat, where xxx is the max amount of virtual address space to be claimed. The value rlp.rlim_cur will then be set to the lower of Vmemory and rlp.rlim_max.

If memory cannot be mapped, the error M00048 "Failed to create memory map for [filename]. Error [detailed message]" will be displayed and put into the errorlog.txt file.

The amount of memory that can be locked

As well as the obvious limitation of physical memory, there is generally a limit set on the amount of memory that can be locked. Another frequently used term for locked is 'wired' – this is the term used in top for example.

On most systems, memory can only be locked by root. Under IBM AIX, memory cannot be locked at all.

Under IRIX, other users can lock memory if the root user has enabled this for a particular application using the `cheap` command. This limit is shown as 'memoryuse' in the `limit` command.

Before a "Failed to lock memory for file xxx" error is given, Mascot Monitor will try and increase the amount of RSS available by calling

```
setrlimit(RLIMIT_RSS, xxx)
```

with the current value plus the size of the file to be locked. Under Solaris, the `RLIMIT_AS` value is used – (rather confusing use of 'AS' by Sun).

Under Linux, prior to RedHat 7.2, the limit of memory that can be locked is hard-coded in the kernel as 50% of all memory. You will need to change the source and rebuild the kernel to overcome this limit.

If the resource limit cannot be increased, then error M00114 "Error calling `setrlimit(RLIMIT_RSS, [memory requested])` - error [detailed error message]" will be put into `errorlog.txt`

If the memory cannot be locked, then the error M00073 "Failed to lock memory for file [file name]. Error [detailed text]" will be put into the `errorlog.txt` file.

If Mascot Monitor, (`ms-monitor.exe`), is a 32 bit executable, the 2Gb limit can quickly be reached by having several large databases locked into memory. To work around this limit, a separate `ms-lockmem.exe` program is provided – this is fork/exec'd from `ms-monitor.exe` when the flag 'SeparateLockMem 1' is added to the options section of `mascot.dat`.

Physical memory

If the amount of memory locked gets close to the amount of physical memory, the system will grind to a halt! The error M00073 "Failed to lock memory for file [file name]. Error [detailed text]" will also probably be put into the `errorlog.txt` file.

Data space

This amount does not include the space used by memory-mapped files.

Insufficient data space will cause a large `master_results.pl` script to fail and a mascot search to fail with an error M00000 – "Out of memory (malloc) [number of] bytes requested"

Swap space

When all physical memory is exhausted, swap space is used. When all swap space is used, no more memory can be allocated and an error will be reported.

There is a different way of setting up swap space on each system – see system documentation.

Mascot shows free swap space for cluster nodes only.

Stack space

Has not been a problem yet.

Thread stack space

The default differs on all platforms; from 5.6K on IBM AIX to 128K on IRIX. However, this is not normally an issue, since it is increased by all the binaries at run time to 128k.

File descriptors

Mascot requires approximately 3 file descriptors per database



Installation: Microsoft Windows

Release Notes

Refer to the release notes for last-minute additions and platform specific notes.

Cluster Mode

If you have a licence to run Mascot on four or more processors, and plan to do so on a networked cluster of machines, then please familiarise yourself with the material in Chapter 11, Cluster Mode, before proceeding with the installation.

Overview

To install Mascot, the following steps need to be performed

1. Verify that the computer has sufficient memory and disk space
2. Verify that the computer has a suitable version of Microsoft Windows installed. Mascot will not run under Windows 9x, Me, XP Home, or Server 2003 for Itanium. Refer to the release notes for Windows Vista support status
3. Virus scanning software or Microsoft Outlook should not be running during the installation
4. Install Web server software if not already installed.
5. Install Perl, if not already installed - supplied on the Mascot CD.
6. Install Mascot

It is essential that steps 4, 5, and 6 are performed in that order

System Requirements

Disk Space

A typical installation of the Microsoft Web server requires about 150 Mb

A typical installation of ActiveState Perl requires about 17Mb

A full installation of Mascot requires approx 1.8 Gb

The hard disk should be formatted for NTFS. FAT32 has a file size limit of 4Gb, which will prevent the use of very large databases

It is strongly advisable that NTFS file compression is *not* used for the compressed database files. There are reports that NTFS compression is not fully compatible with memory mapping. NTFS file compression can be used on the FASTA and reference files if you wish.

Memory

To get the best performance from Mascot, the database files need to be memory mapped. The installation program will try to lock the supplied database, SwissProt, into memory if the computer has sufficient RAM. If successful, this reduces the free RAM, available for other applications, by approximately 120 MB. It is recommended that you have at least 1 GB of RAM, and preferably 2 GB.

Microsoft Windows versions

NT 4 SP6a

Microsoft stopped supporting Windows NT Workstation 4.0 and Windows NT Server 4.0 in 2004. This will be the final release of Mascot to support NT4. We recommend that you upgrade to one of the more recent versions of Windows, (XP, Server 2003, or Vista).

Service pack 6a is required:

<http://microsoft.com/ntserver/nts/downloads/>

Mascot 2.2 uses a new installer, so as to support 64 bit versions Windows. This installer is not compatible with NT4, so you will need to request a special CD that contains the old installation program. See the release notes for further details.

The web server for NT *Server* (on the CD labelled Option Pack 4) is Microsoft Internet Information Server 4 (IIS 4). The same CD can be used to install IIS on Windows NT *Workstation*, but the product is then called Microsoft 'Personal Web Server' (PWS). Option Pack 4 is not

supplied with NT Workstation, but is available for download as Option Pack x86 at:

<http://microsoft.com/ntserver/nts/downloads/>

If there is a copy of the Option pack CD anywhere *within your organisation*, then Microsoft grants you the rights to install from that CD.

When installing Option Pack 4, all the required components for Mascot are included in the 'Typical' option. **Do not choose the 'Minimum' option.** If you choose the 'Custom' option, do not de-select any of the default components.

If you are installing from the Internet, or upgrading from an earlier version of IIS (the latter is a common when installing onto NT server), you should not select the 'Upgrade Only' option, but choose to also add components. Do not de-select any of the default components.

2000 Professional and Server

Mascot will run under Microsoft Windows 2000 Professional and Server. In general, 2000 Professional is sufficient. The only significant limitation is that the number of simultaneous connections is limited to 10 with Microsoft's IIS web Server. These limits do not apply to 2000 Server.

Service pack 4 should be installed:

<http://www.microsoft.com/windows2000/downloads/>

The Microsoft web server for Windows XP is IIS 5.0, which is provided as part of the standard distribution. If IIS is not installed, choose 'Add or Remove Programs' in the Control Panel. Select 'Add/Remove Windows Components', and check the box for Internet Information Services.

XP

Mascot will run under Windows XP Professional; Windows XP Home is not supported.

The Microsoft web server for 32 bit editions of Windows XP is IIS 5.1, which is provided as part of the standard distribution. If IIS is not installed, choose 'Add or Remove Programs' in the Control Panel. Select 'Add/Remove Windows Components', and check the box for Internet Information Services. XP Professional x64 Edition uses IIS 6.0, the same as Server 2003.

It is advisable to ensure that the latest service pack has been installed. Check the following URL for current information:

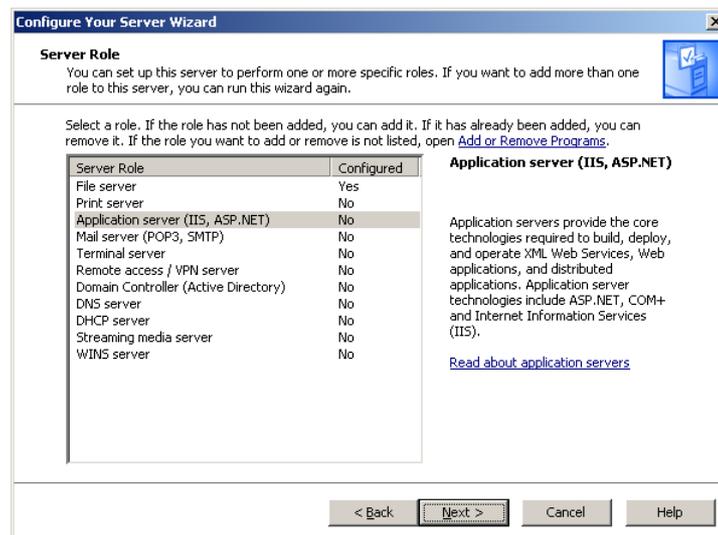
<http://www.microsoft.com/windowsxp/downloads/>

Server 2003

Mascot will run under all editions of Windows Server 2003 except those for the Itanium processor.

The Microsoft web server for Windows XP is IIS 6.0

You may need to install IIS by configuring the server as an Application Server. When you start your server, you should see a 'Manage Your Server' wizard. If not, go to Administrative Tools and click on 'Manage Your Server'. When the wizard opens, click on 'Add or Remove a Role', then select Application Server



Proceed through the wizard, accepting all the defaults, to install IIS.

It is advisable to ensure that the latest service pack has been installed. Check the following URL for current information:

<http://www.windowsupdate.com/>

Vista

Refer to the release notes for Vista support status.

The Microsoft web server for Vista is IIS 7.0, which is provided as part of the standard distribution. Refer to the release notes for installation details.

It is advisable to ensure that the latest service pack has been installed. Check the following URL for current information:

<http://www.windowsupdate.com/>

Web Server

Mascot is compatible with most web servers, and we do not recommend any specific product. However, Microsoft has ensured that its own server products are closely integrated with the various Windows platforms, and the procedures for their installation and configuration follow familiar Windows conventions. The Mascot installation has been fully automated for Microsoft Internet Information Server 4.0 and later.

A good starting point for IIS support information is

<http://www.iis.net/>

If you wish to use a different Web server, such as Apache, you will need to perform some manual configuration. Appendix D provides configuration information for some of the more widely used web servers.

Perl

Perl is the scripting language used by Mascot. Version 5.6 or later is required.

Active Perl 5.8.8 from ActiveState Corporation is supplied on the Mascot CD. You must install Perl from the CD *after* installing the Web server, and *before* installing Mascot.

To install ActivePerl from the CD, in Windows Explorer, double click on the file:

ActivePerl-5.8.8.820-MSWin32-x86-274739.msi

It is recommended that you accept all the default options for the installation. If you prefer, you may be able to download a later version of ActivePerl from the ActiveState web site:

<http://www.activestate.com/store/activeperl/>

Note: If you are using Windows NT4 or an early version of Windows 2000, you may need to update the Windows Installer before you can install ActivePerl. In Windows Explorer, double click on the file *InstMsiW.exe*, which can be found in the *unpacked\html\downloads* directory on the Mascot CD.

Note: If you have to re-install IIS, you should re-install Perl.

Any problems that you might encounter are almost certainly addressed in the FAQ's that form part of the ActivePerl product documentation.

ActiveState Marketing Requirements

The following statements are included to comply with the ActiveState Redistribution Agreement:

Commercial support for ActivePerl is available through ActiveState at:

<http://www.activestate.com/Support/Enterprise/>

For peer support resources for ActivePerl issues see:

<http://www.activestate.com/Support/>

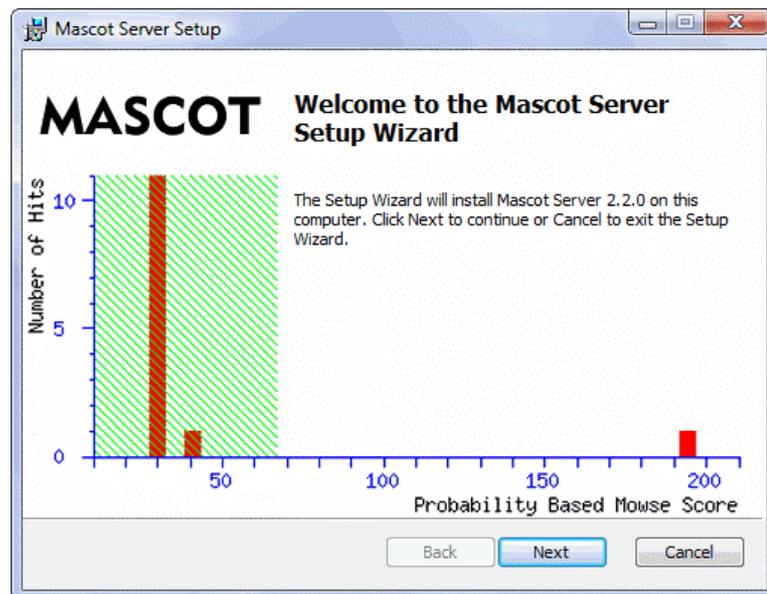
The ActiveState Repository has a large collection of modules and extensions in binary packages that are easy to install and use. To view and install these packages, use the Perl Package Manager (PPM) which is included with ActivePerl.

ActivePerl is the up-to-date, quality-assured ActivePerl binary distribution from ActiveState. Current releases and other professional tools for open source language developers are available at <http://www.activestate.com>

Mascot Installation

From 'My Computer' or Windows Explorer, double click on the *Setup.exe* program on the Mascot CD

The following window will be displayed:



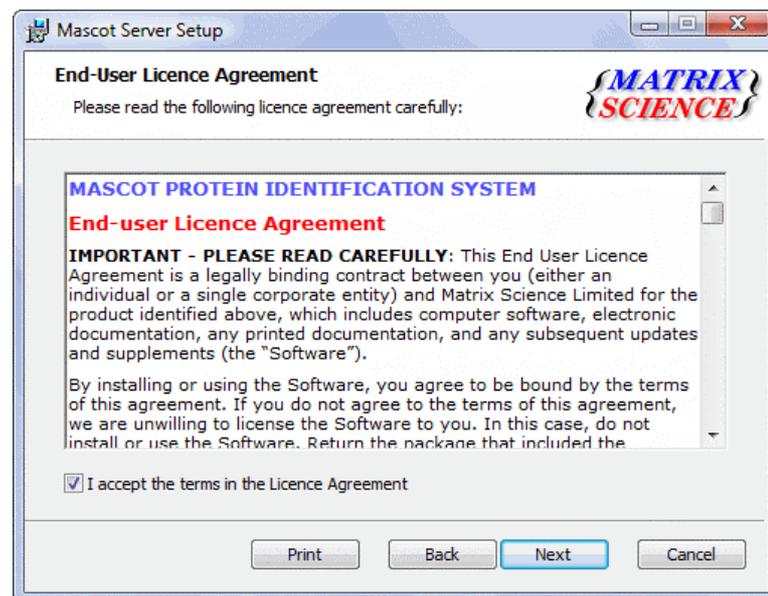
If the installation cannot proceed, a message box will be displayed. Typical problems include:

You do not have Administrator privileges: Log out and log in as a user with local Administrator privileges

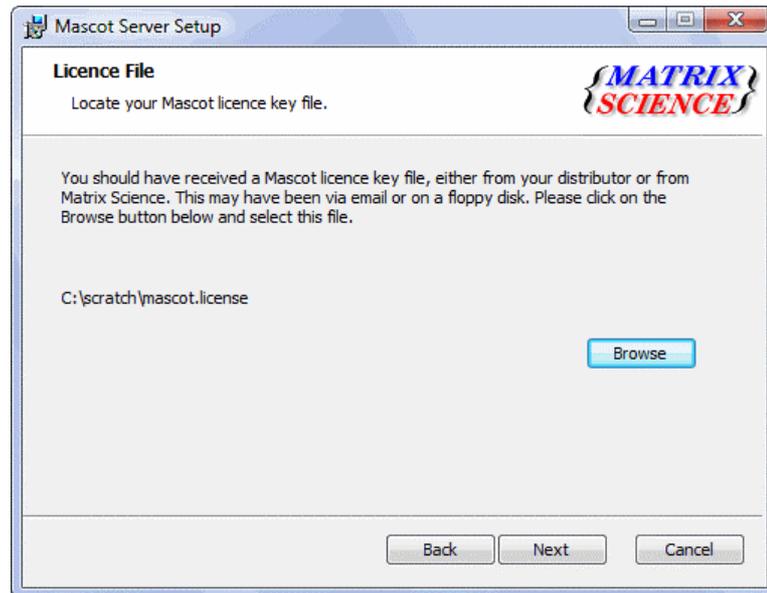
Perl is not installed: Install Perl as described above

Unsupported Windows platform: Refer to the system requirements at the beginning of this Chapter

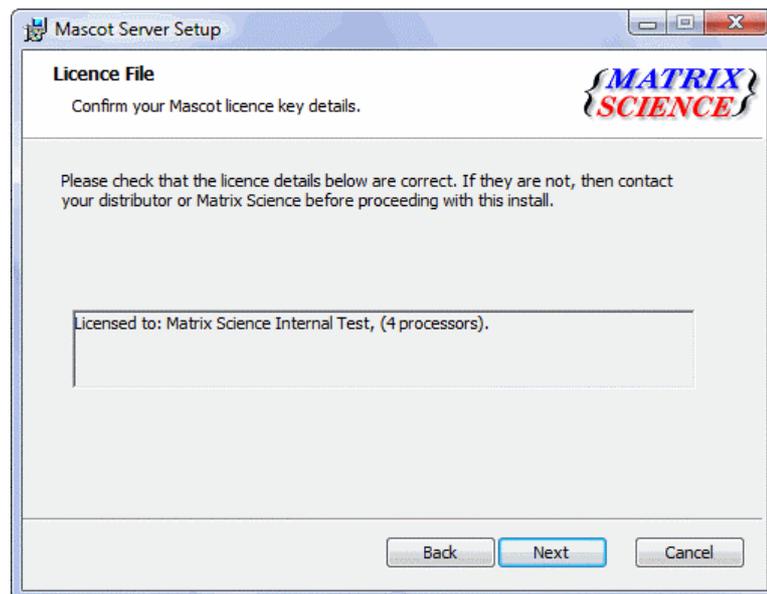
Any problem(s) must be fixed before the installer will proceed. Pressing Next displays the Mascot End-User Licence Agreement:



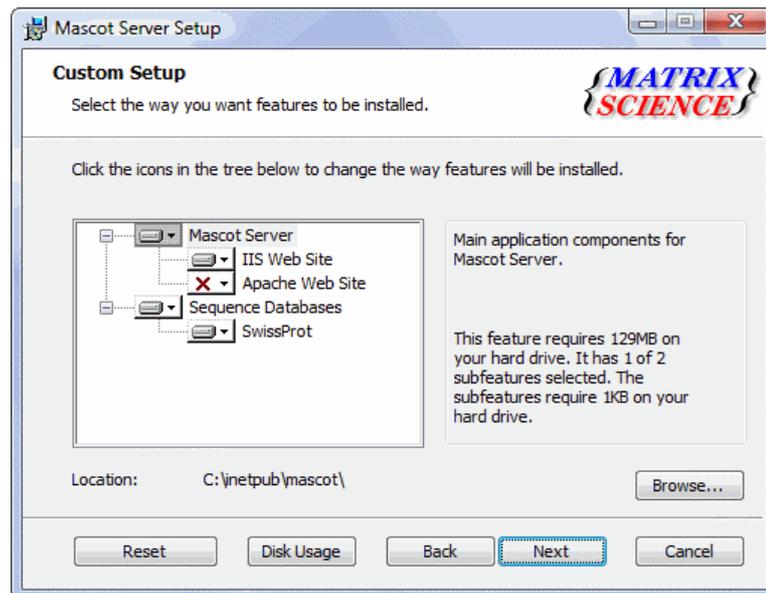
You must consent to proceed with the installation. Press Next and you will be asked to locate your licence file. This is a text file that has been sent by email or supplied on removable media. NB The licence file must be called *mascot.license*.



The next screen confirms that a valid licence file has been located



You cannot proceed without a valid licence. The next screen allows you to choose which components will be installed:

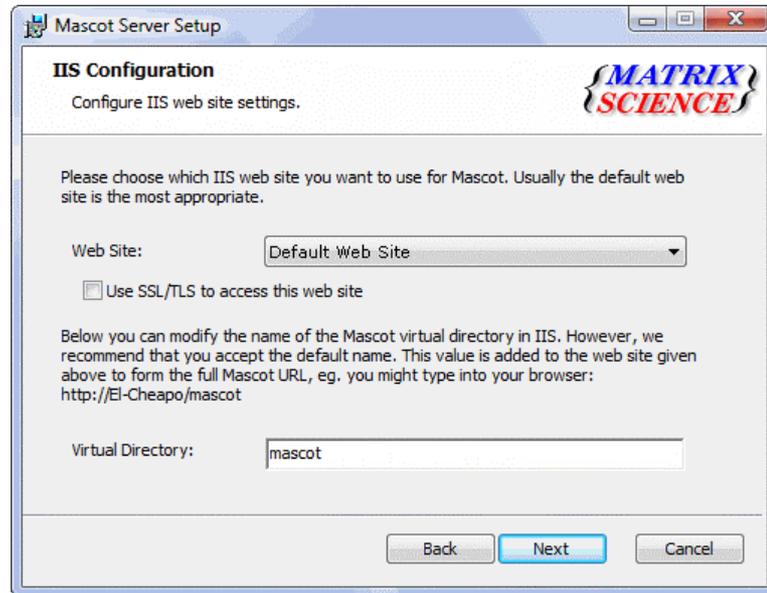


If IIS is installed and functional, the default choices are as shown above, with IIS being configured automatically. If you don't have IIS installed, the Apache option will be selected instead. A test for whether Apache or some other web server is actually installed comes later.

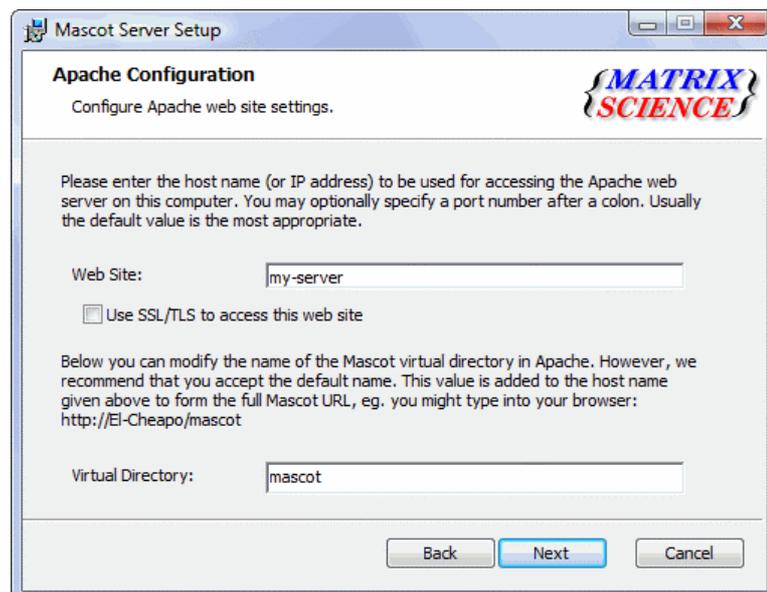
You can de-select the Swiss-Prot database, but if this is a clean install, you are advised not to do so. It is better to proceed with a full installation, so that Mascot can be shown to be correctly installed and functional. If you don't want SwissProt to be available, you can easily remove it later.

The default location for the installation is `C:\inetpub\mascot` with the sequence databases in `C:\inetpub\mascot\sequence`, but you can change one or both of these by selecting the component then choosing Browse. If there is insufficient disk space on the selected drive(s), the installation will not be able to continue.

The next step depends on whether IIS or Apache was selected as the web server. For IIS, there will be a drop-down list of all the available web sites. In most cases, you should choose 'Default Web Site'. If you choose a different web site, refer to the notes on multiple web sites in the 'Manual Configuration' section of this chapter.



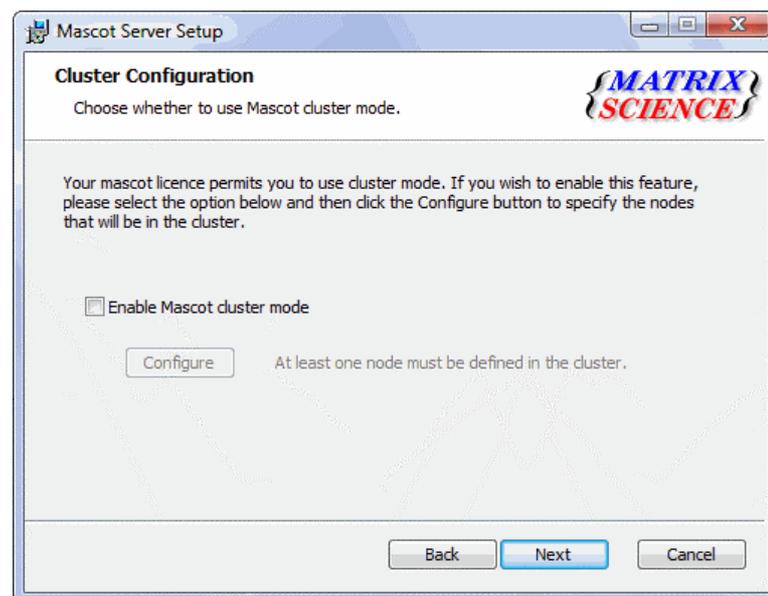
For Apache, or any other web server, you need to confirm the local web server hostname and port. NB Do not enter localhost as the web site if you wish to access your Mascot server from other computers on your LAN. If there are DNS problems, so that a hostname is not recognised across the LAN, then enter an IP address. The default ports are 80 for http and 443 for https. The installer will test that the web server responds using the specified hostname and port number.



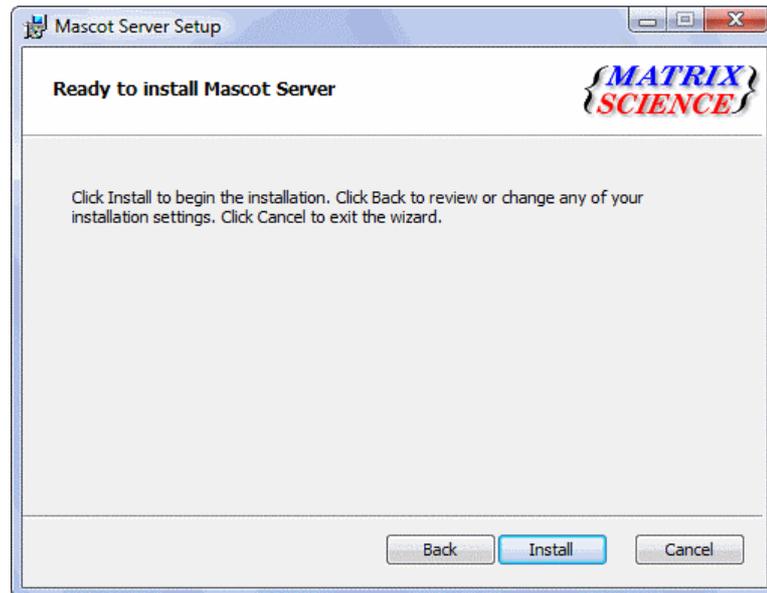
If you have configured your web server as a secure server (https), check the box for 'Use SSL/TLS to access this site'.

The virtual directory name can be changed if you wish, but remember that users are more likely to guess the correct URL if you stick with mascot.

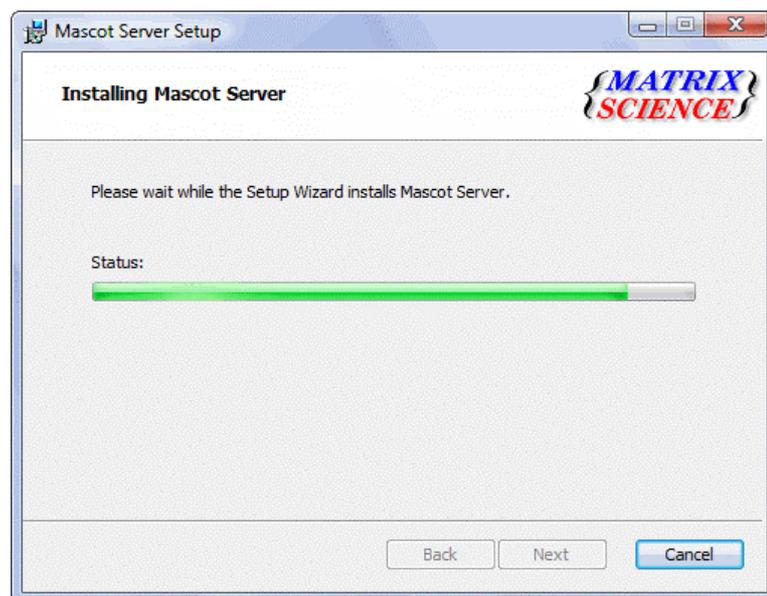
You will only see the following screen if your Mascot licence is for 4 or more processors and this is a first time installation. This dialog is used to configure Mascot for execution on a networked cluster. If you intend installing Mascot on cluster, refer to Chapter 11 for further details before proceeding. If you are installing Mascot on a single multiprocessor server, leave the Enable cluster mode checkbox clear.



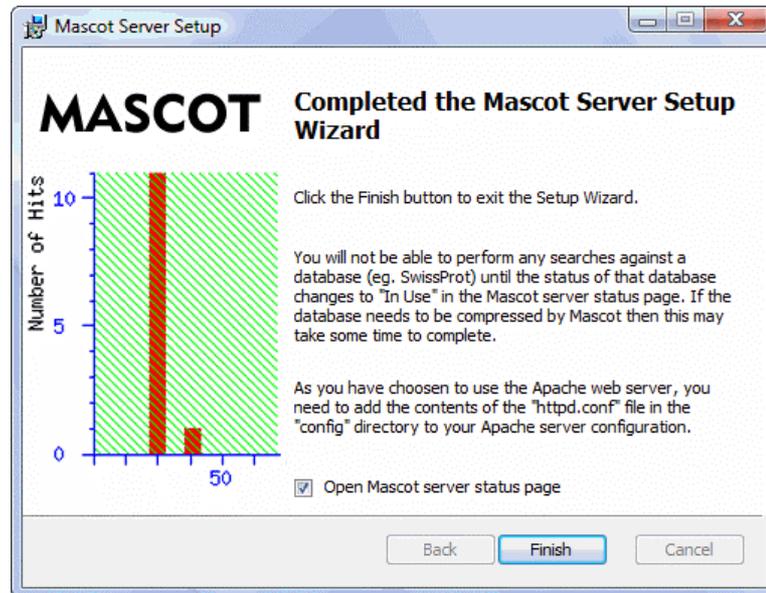
The next step is your last opportunity to cancel the installation!



Copying the files takes only a few minutes

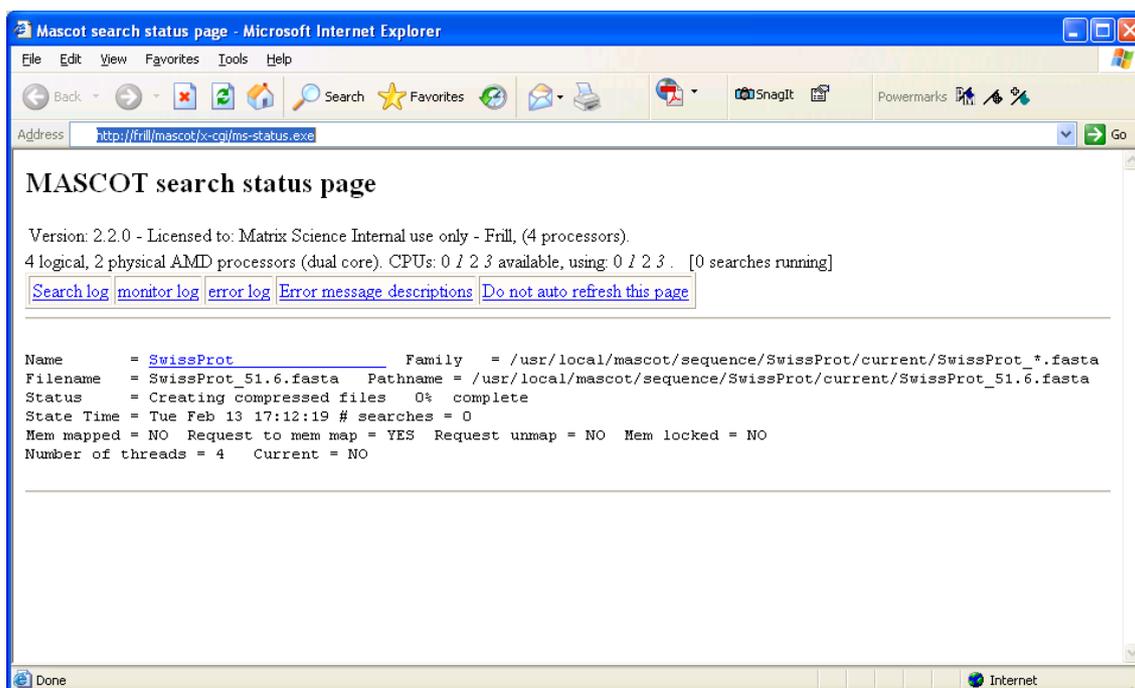


The final step confirms that the installation was successful.



If you are using Apache, refer to Appendix D for configuration instructions.

Although the installation is now complete, there will be a lot of disk activity as the SwissProt database is compressed and a test search is run to verify system operation. Searches cannot be performed until the test search has been completed successfully. Unless you clear the checkbox, when you press Finish, the database status screen will be displayed in the default web browser:



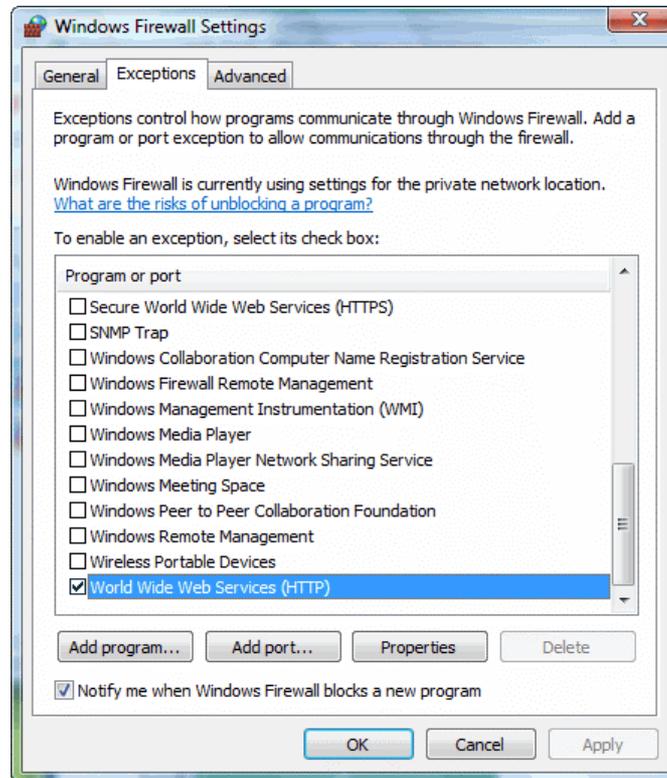
If all is well, the status line will display the following series of messages:

```
Creating compressed files
Running 1st test
First test just run OK
Trying to memory map files
Just enabled memory mapping
In Use
```

You can begin exploring and using Mascot. However, do not try to run searches or view results reports until the relevant sequence database is 'In Use'.

Windows Firewall

If Windows Firewall is enabled, you may be blocked from accessing the Mascot server from other computers. If so, you need to open up port 80. From the Control Panel, choose Windows Firewall. On the exceptions tab, check the box for 'World Wide Web Services (HTTP)'.



Security

Mascot security is disabled on installation. To enable Mascot security, refer to Chapter 12

Manual Configuration

LCQ_DTA

This script, available from the main Mascot menu, makes it possible to upload a Finnigan LCQ or LTQ *.raw file for searching by Mascot. The html form executes *lcq_dta.exe* or *extract_msn.exe*, which are licensed software, copyright University of Washington, and normally supplied with LCQ or LTQ instruments as part of the Xcalibur data system. The script supplied with Mascot expects to find the executable in the path `C:\LCQ\system\programs\lcq_dta.exe`. If the program is in another directory, or if using *extract_msn.exe*, then open the file `C:\InetPub\Mascot\cgi\lcq_dta_shell.pl` in an editor such as Notepad, and modify the following line:

```
my $lcqExe = "c:\\LCQ\\system\\programs\\lcq_dta.exe";
```

Note that the backslashes used as directory delimiters must be entered in pairs, as shown above.

The script needs to create temporary files, and it uses a directory `C:\TEMP`. If this does not exist, you should either create it, or change the following line to point to a suitable temporary directory.

```
my $tempDir = "c:\\temp";
```

To use the `lcq_dta` form, enter the filename and any other parameters required, and press the "Generate .DTA Files" button. After a few seconds, the Mascot search screen will be displayed. Enter search parameters and proceed as normal.

User help can be found by choosing Help from the Mascot main menu and then 'Thermo Finnigan Xcalibur'.

GD

The GD module is a Perl interface to Thomas Boutell's `libgd` library. It can be downloaded from any CPAN site:

```
http://search.cpan.org/dist/GD/
```

The Mascot installation program automatically installs GD. If you later change the version of Perl, you may need to re-install GD using the Perl Package Manager. In early versions of ActivePerl, this is a perl script called `ppm.pl`. In later versions, it is a DOS batch file, `ppm.bat`:

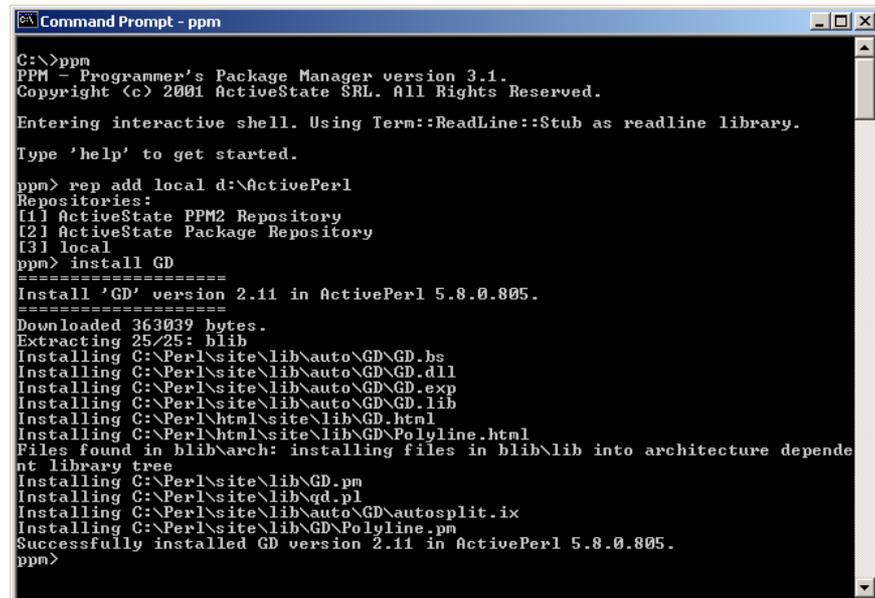
For Perl 5.6, you can install precompiled packages for GD using the ActiveState Perl Package Manager (PPM). At a command prompt, execute the following commands:

```
ppm
install GD
quit
```

For Perl 5.6 or 5.8, GD can be installed from the Mascot CD-ROM. At a command prompt, execute the following commands:

```
ppm
rep add local d:\ActivePerl
install GD
quit
```

Where *local* is an arbitrary name used to identify the package repository, and *d*: is the drive letter for the Mascot CD-ROM.



```

C:\>ppm
PPM - Programmer's Package Manager version 3.1.
Copyright (c) 2001 ActiveState SRL. All Rights Reserved.

Entering interactive shell. Using Term::ReadLine::Stub as readline library.
Type 'help' to get started.

ppm> rep add local d:\ActivePerl
Repositories:
[1] ActiveState PPM2 Repository
[2] ActiveState Package Repository
[3] local
ppm> install GD
=====
Install 'GD' version 2.11 in ActivePerl 5.8.0.805.
=====
Downloaded 363039 bytes.
Extracting 25/25: blib
Installing C:\Perl\site\lib\auto\GD\GD.bs
Installing C:\Perl\site\lib\auto\GD\GD.dll
Installing C:\Perl\site\lib\auto\GD\GD.exp
Installing C:\Perl\site\lib\auto\GD\GD.lib
Installing C:\Perl\html\site\lib\GD.html
Installing C:\Perl\html\site\lib\GD\Polyline.html
Files found in blib\arch: installing files in blib\lib into architecture depende
nt library tree
Installing C:\Perl\site\lib\GD.pm
Installing C:\Perl\site\lib\gd.pl
Installing C:\Perl\site\lib\auto\GD\autosplit.ix
Installing C:\Perl\site\lib\GD\Polyline.pm
Successfully installed GD version 2.11 in ActivePerl 5.8.0.805.
ppm>

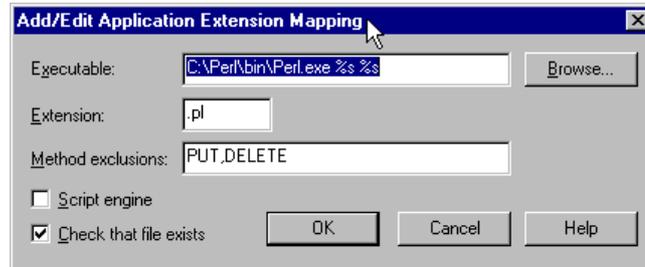
```

Multiple web sites

Using IIS on Server versions of Windows, it is possible to create multiple web sites. If multiple web sites exist when Mascot is installed, you will be offered the choice of which web site to install under. However, there are a number of issues to be aware of if the “Default Web Site” is not used.

Perl installation.

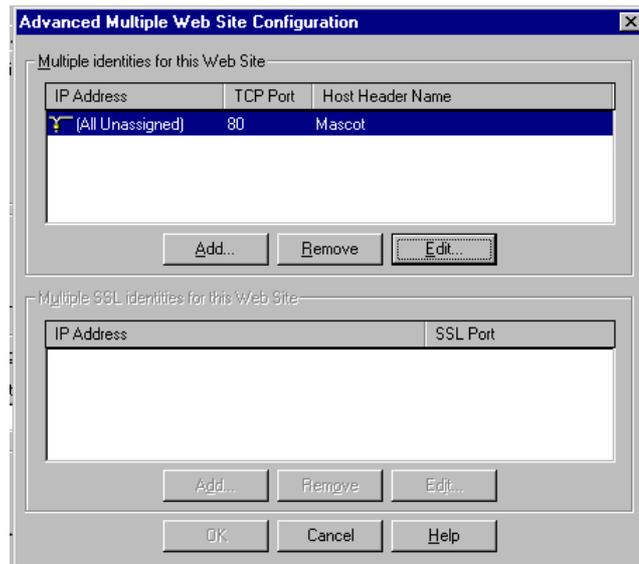
On installation, ActiveState Perl only sets up mappings for the default web site. To add mappings for another web site, right click on the new web site, choose properties, and then click on the “Home Directory” tab. Then, click on the configuration button. If there is no entry in the list for .pl extension, click on “Add” and enter the following information:



Host name

Mascot will be installed in the correct virtual directory, but the host (base) name may be wrong – the installation program has chosen the computer name.

If you have set up multiple web sites, then it is probable that you have created a DNS entry that is the same as the “Host header Name”. In this case, replace the computer name with the “Host header name”. Refer to the IIS documentation for details on setting up multiple web sites using the same IP and port address. Briefly, you will need to add a new web site, and then click on the “Web site tab” of the properties box. Next, click on the Advanced button, and enter a host name:



To memory lock databases totalling more than 2GB

For 32 bit editions of Windows, there is a 2 GB limit on the address space for any single process. Mascot Monitor, (ms-monitor.exe), can easily reach this limit by trying to lock several large databases into memory. To work around the 2GB limit, a separate ms-lockmem.exe program is provided – this is fork/exec'd from ms-monitor.exe when the flag 'SeparateLockMem 1' is added to the options section of mascot.dat. Further details can be found in Chapter 7.

Hyper-threading

Hyper-threading is a technique used by Intel to improve the performance of multi-threaded programs by putting multiple 'processors' onto the same physical chip. The performance does not increase linearly with the number processors, (currently limited to 2), because the individual processor cores share other resources such as the on-chip cache. In many cases, there is little performance improvement for Mascot searches when hyper-threading is enabled. On most systems, a BIOS setting can be used to enable and disable hyper-threading.

Hyper-threading is detected automatically and the additional processors may be used to increase performance. For example, with a 2 CPU Mascot license on a dual processor Xeon system, there will appear to be 4 processors. On this system, the Mascot search status screen will display:

```
4 processors available, [Hyper-threading enabled], (numbers 0 1 2 3 ).
Using processor numbers 0 1 2 3.
```

(Processor numbers 2 and 3 are in italics to show that they are the 'hyper-threaded' cpus)

On the same system, with a single processor license, the Mascot search status screen will display:

```
4 processors available, [Hyper-threading enabled], (numbers 0 1 2 3 ).
Using processor numbers 0 2.
```

Processor numbers 0 and 2 are chosen by default because they are on the same physical chip. Attempting to set Processors=0,1 in the processors section of mascot.dat will fail.

To make use of these other processors, use the database maintenance utility to change the number of threads to the number of logical processors. For example, on a dual CPU Xeon system, change the number from 2 to 4.

Under Windows, it is likely that the greatest performance increase will be seen when running Windows 2003 server. With NT4, and in some cases Windows 2000, performance may actually be worse with hyper-threading enabled.

Troubleshooting

Check the Support pages on the web

There may be a fix listed on the Matrix Science Web Site. From the menu, choose Support, and scan down to see if your problem is described.

Log Files Created During Installation

There are two files created during the installation that may help you understand and correct problems during installation. You may also be asked send these files to the person providing technical support.

Once all the files have been copied from the CD-ROM, a file named *install.log* will be created in the mascot logs directory (normally *C:\InetPub\Mascot\logs*). This file has information about the configuration of your system, and any problems found during installation and configuration. You should check this file first.

If the installation program 'hangs' without completing, *install.log* will be found in your temp directory. This is normally something like *C:\Documents and Settings\myname\Local Settings\Temp*.

The installation program doesn't recognise Perl

To test whether Perl is really installed, you should open a command prompt, and type:

```
perl -v
```

The version number should be displayed. If this seems to be functioning correctly, please re-start the computer and then re-run the Mascot installation program. If this still fails, you may continue with the Mascot installation, but you will probably need to install the GD library manually (see above).

If, when you type **perl -v** you see the text:

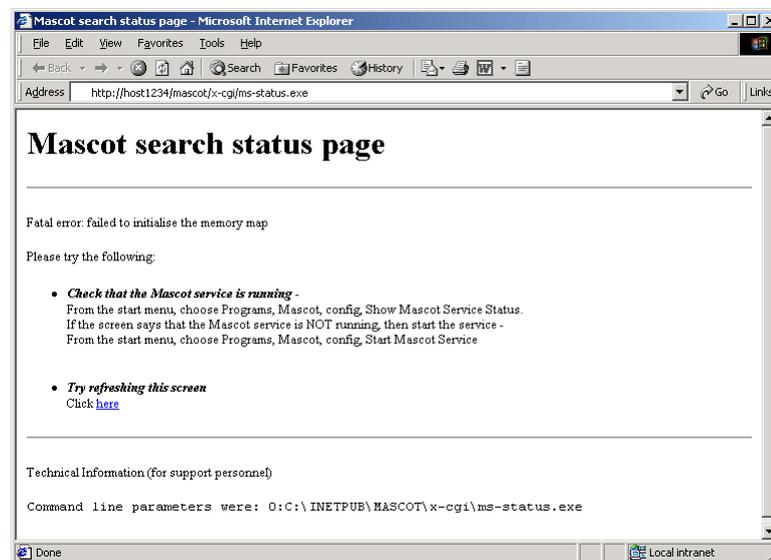
```
The name specified is not recognized as an  
internal or external command, operable program or  
batch file.
```

then Perl is not installed or is not on the path. If you have just installed it, you should try restarting the computer and performing the test again.

If that fails, try re-installing Perl, making sure that you choose the option to put it onto the path.

The status screen shows an error

If the Mascot Monitor service fails to start, then the following text or something similar will be displayed in the status screen:



There are several possible causes:

Service not started

Since one of the first things that the Monitor service does is to create the memory mapped file, this could indicate that the service has not started. You can tell whether the service has started, by running the program from the Start menu, choose *Start; Programs; mascot; config; Show Mascot ms-monitor service status*.

If the service is not running, check the *monitor.log* and *errorlog.txt* file in the *logs* directory. If there is nothing in those files, then it may be necessary to try and run *ms-monitor.exe* as a command line executable. *You should only do this if the Mascot service is not running*. To do this, open a command prompt window, and change directory to the mascot *bin* directory. If your installation path was the default, you will need to type:

```
cd \Inetpub\mascot\bin
```

next start the monitor program:

ms-monitor DEBUG

Any error messages should be displayed on the screen. If possible, correct the faults, and then start the Mascot Service from the start menu. Note that the mascot service should never be running at the same time as *ms-monitor.exe* is being run from the command line.

URL mappings

If the service is running, then there may have been a problem with the URL mappings. This could be caused by trying to have more than one version of Mascot installed. Does the error message indicate that the wrong *ms-status.exe* is being run? If so, you will need to change the settings manually.

International Versions Of Windows

If Mascot is installed on a version of Windows that is not in the English language, then when the *ms-status* screen is displayed, it may have the error ‘Failed to initialise memory map’

To correct this fault, the following procedure is required:

1. You will need to find the names of the ‘groups’ that your version of NT uses for Administrators and Guests. In German, for example, these names are “Administratoren” and “Gaste” respectively. To see a list of User names, from the start menu, select Programs, Administrative Tools (common), User Manager. The section at the bottom of the screen displays the group names. Make a note of the two names.
2. From the start menu, select
Programs | Mascot | Config | Stop Mascot Service
3. From the start menu, select
Programs | Mascot | Config | Mascot Configuration File
4. Scroll down to near the bottom of the file and find the line:
NTIUserGroup Guests
and change this to (for example, for German)
NTIUserGroup Gaste
5. Find the line
NTMonitorGroup Administrators
and change this to (for example, for German)
NTMonitorGroup Administratoren
6. Save the *mascot.dat* file
7. Delete the files:
c:\inetpub\mascot\sequence\MSDB\current\MSDB*.a00

c:\inetpub\mascot\data\mascot.control

(Note that these files may be in a different directory if you did not install mascot in the default location)

8. From the start menu, select
Programs | Mascot | Config | Start Mascot Service
9. Re-load the status page:
Programs | Mascot | Search Status
(You may need to re-fresh / re-load the page)

Wait until the files have been compressed and a test search has been done. Mascot is now ready for use.

The local search facility (online help index) does not work

The online help pages are indexed using a product called ht://Dig. A log file is made as the indexes are built during the installation. The log file *mascot\htdig\build.log* may contain an error message indicating the nature of the problem.

Search status shows a failure to create compressed files

On the search screen, find out what caused the error by clicking on the *Error log* link, fix the fault, (possibly out of disk space), and then click on *retry*.

4

Validation

CGI Operation

To verify that the search engine is functioning correctly when executed as a CGI application, launch a JavaScript aware web browser and load the Mascot home page, (http://your_server/mascot/). Select Mascot from the main menu and then choose the “Peptide Mass Fingerprint” link near the top of the page. This will load the search form for a peptide mass fingerprint.

Enter your name and email address into the fields at the top of the form and type a number, say 1234, into the Query field. Then press the Start Search... button.

The search form will be replaced by the search progress screen. This has a few lines of text at the top, ending in the line “Searching ...”. Additional lines will appear showing the percentage of the search that has been completed.

Once the search is complete, the Master Results page will appear. Unless you went to the trouble of entering some real mass values, the results will be meaningless!

Monitor Test

When Mascot Monitor is started, it runs a test search against each sequence database. It also runs this same test search against any update to the database as part of the exchange procedure. If the test search fails, an error message will be displayed in the Mascot Status screen and the database will not be available for searching. Error messages from Monitor are logged to *errorlog.txt* in the *mascot/logs* directory. Both this file and *monitor.log* can be viewed using links on the Mascot Status page.

The input file which defines a test search can be found in the *mascot/data/test* directory. The filename is constructed from the name of the

database together with the extension `.asc`. For example,
`SwissProt.asc`.

Note: Test files for new databases are generated by modifying `OWL.asc`.
Never delete this file.

The output of the test search may change slightly with each new update
to a database. Sequences may be corrected or descriptions modified.
Quite often, a new entry appears which is very homologous with one of
the matched proteins so that it appears on the hit list.

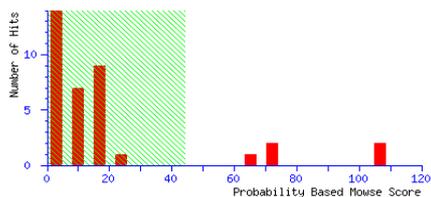
Using SwissProt 51.6, the report from running the standard test search,
`SwissProt.asc`, is shown on the following pages.

(MATRIX)
(SCIENCE) Mascot Search Results

User : Monitor Test DB 0
 Email :
 Search title :
 MS data file : dyckall.asc
 Database : SwissProt 51.6 (257964 sequences; 93947433 residues)
 Timestamp : 18 Feb 2007 at 19:03:19 GMT
 Protein hits : [HSP7D_MANSE](#) Heat shock 70 kDa protein cognate 4 (Hsc 70-4) - Manduca sexta (Tobacco hawkmoth) (Tobacco hornworm)
[HSP7D_ECHGR](#) Heat shock cognate 70 kDa protein (HSP70) - Echinococcus granulosus
[HSP7A_DROSI](#) Heat shock 70 kDa protein cognate 1 (Heat shock 70 kDa protein 70C) (Fragments) - Drosophila simula
[HSP7L_HUMAN](#) Heat shock 70 kDa protein 1L (Heat shock 70 kDa protein 1-like) (Heat shock 70 kDa protein 1-Hom) (
[HSP7D_DROME](#) Heat shock 70 kDa protein cognate 4 (Heat shock 70 kDa protein 88E) - Drosophila melanogaster (Fru

Probability Based Mouse Score

Ions score is $-10 \cdot \log(P)$, where P is the probability that the observed match is a random event. Individual ions scores > 44 indicate identity or extensive homology ($p < 0.05$). Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.



Peptide Summary Report

Format As Peptide Summary [Help](#)

Significance threshold $p < 0.05$ Max. number of hits AUTO

Standard scoring MudPIT scoring Ions score or expect cut-off 0 Show sub-sets 0

Show pop-ups Suppress pop-ups Sort unassigned Decreasing Score Require bold red

[Import results into MI](#)

Select All Select None Search Selected Error tolerant Archive Report

1. [HSP7D_MANSE](#) Mass: 78630 Score: 107 Queries matched: 2
 Heat shock 70 kDa protein cognate 4 (Hsc 70-4) - Manduca sexta (Tobacco hawkmoth) (Tobacco hornworm)
 Check to include this hit in error tolerant search or archive report
- | Query | Observed | Mr(expt) | Mr(calc) | % | Miss Score | Expect Rank | Peptide |
|---------------------------------------|-----------|------------|------------|----------|------------|-------------|---|
| <input checked="" type="checkbox"/> 1 | 671.90000 | 1341.78545 | 1341.72522 | 0.00449 | 0 | 63 0.00081 | 1 K.DAGTISGLNVLR.I |
| <input checked="" type="checkbox"/> 3 | 973.90000 | 1945.78545 | 1945.92036 | -0.00693 | 0 | 46 0.038 | 1 K.NQVA N PNHTIFDAK.R + Oxidation (M) |
2. [HSP7D_ECHGR](#) Mass: 79244 Score: 104 Queries matched: 2
 Heat shock cognate 70 kDa protein (HSP70) - Echinococcus granulosus
 Check to include this hit in error tolerant search or archive report
- | Query | Observed | Mr(expt) | Mr(calc) | % | Miss Score | Expect Rank | Peptide |
|---------------------------------------|-----------|------------|------------|---------|------------|-------------|---------------------|
| <input type="checkbox"/> 1 | 671.90000 | 1341.78545 | 1341.72522 | 0.00449 | 0 | 63 0.00081 | 1 K.DAGTISGLNVLR.I |
| <input checked="" type="checkbox"/> 2 | 808.30000 | 1614.58545 | 1613.75732 | 0.05132 | 0 | 41 0.099 | 1 R.TTPSYVAFTDTER.L |

Proteins matching the same set of peptides:

52 Mascot: Installation and Setup

[HSP7C_ORYLA](#) Mass: 83745 Score: 102 Queries matched: 2
Heat shock cognate 71 kDa protein (Hsc70.1) - *Oryzias latipes* (Medaka fish) (Japanese ricefish)

[HSP71_PUCGR](#) Mass: 76383 Score: 102 Queries matched: 2
Heat shock protein HSP1 - *Puccinia graminis*

3. [HSP7A_DROSI](#) Mass: 25511 Score: 74 Queries matched: 2
Heat shock 70 kDa protein cognate 1 (Heat shock 70 kDa protein 70C) (Fragments) - *Drosophila simula*

Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	% Miss	Score	Expect	Rank	Peptide
2	808.30000	1614.58545	1613.75731	0.05132	0	29	1.6	2 R.TTPSYVAFTESER.L
3	973.90000	1945.78545	1945.92036	-0.00693	0	46	0.038	1 K.NQVAHNPNTIFDAK.R + Oxidation (M)

Proteins matching the same set of peptides:
[HSP7A_DROME](#) Mass: 74963 Score: 73 Queries matched: 2
Heat shock 70 kDa protein cognate 1 (Heat shock 70 kDa protein 70C) - *Drosophila melanogaster* (Fru)

4. [HSP70L_HUMAN](#) Mass: 76939 Score: 72 Queries matched: 2
Heat shock 70 kDa protein 1L (Heat shock 70 kDa protein 1-like) (Heat shock 70 kDa protein 1-Hom) (

Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	% Miss	Score	Expect	Rank	Peptide
2	808.30000	1614.58545	1613.75732	0.05132	0	41	0.099	1 R.TTPSYVAFDTER.L
3	973.90000	1945.78545	1945.92038	-0.00693	0	31	1.1	2 K.NQVAHNPQNTVFDK.R + Oxidation (M)

Proteins matching the same set of peptides:
[HSP70L_MOUSE](#) Mass: 77328 Score: 72 Queries matched: 2
Heat shock 70 kDa protein 1L (Heat shock 70 kDa protein 1-like) (Heat shock 70 kDa-like protein 1)
[HSP70L_RAT](#) Mass: 77240 Score: 72 Queries matched: 2
Heat shock 70 kDa protein 1L (Heat shock 70 kDa protein 1-like) (Heat shock 70 kDa protein 3) (HSP7
[HSP70_SOYBN](#) Mass: 78078 Score: 72 Queries matched: 2
Heat shock 70 kDa protein - Glycine max (Soybean)
[HSP70_XENLA](#) Mass: 77479 Score: 70 Queries matched: 2
Heat shock 70 kDa protein (HSP70) - *Xenopus laevis* (African clawed frog)

5. [HSP7D_DROME](#) Mass: 77695 Score: 63 Queries matched: 2
Heat shock 70 kDa protein cognate 4 (Heat shock 70 kDa protein 88E) - *Drosophila melanogaster* (Fru)

Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	% Miss	Score	Expect	Rank	Peptide
2	808.30000	1614.58545	1613.75732	0.05132	0	41	0.099	1 R.TTPSYVAFDTER.L
3	973.90000	1945.78545	1946.94078	-0.05934	0	22	9.4	3 K.NQVAHNPQTQIFDAK.R + Oxidation (M)

Peptide matches not assigned to protein hits: (no details means no match)

Query	Observed	Mr(expt)	Mr(calc)	% Miss	Score	Expect	Rank	Peptide
<input checked="" type="checkbox"/> 4	1084.90000	2167.78545	2168.13689	-0.01621	0	5	3.5e+02	1 HJHGMALLVPLVYTIK + 3 Oxidation (M)

Search Parameters

Type of search : MS/MS Ion Search
Enzyme : Trypsin
Fixed modifications : SMA (K), SMA (N-term)
Variable modifications : Oxidation (M)
Mass values : Monoisotopic
Protein Mass : Unrestricted
Peptide Mass Tolerance : ± 0.1 %
Fragment Mass Tolerance : ± 0.5 Da
Max Missed Cleavages : 1
Number of queries : 4

Mascot: <http://www.matrixscience.com/>

5

Sequence Database Setup

Database names, URL's, and even formats, change constantly. The most up-to-date information, including configuration screenshots for many popular databases, can be found on the Matrix Science web site at http://www.matrixscience.com/help/seq_db_setup.html

This chapter will guide you through configuring Mascot to search a new sequence database. The examples used will be the databases supplied on the Mascot Databases DVD.

The FASTA Format

Mascot can search any FASTA format sequence database. The FASTA format is extremely simple. Each entry consists of a one line title followed by one or more lines containing the contiguous sequence string in 1 letter code. FASTA databases can contain either amino acid sequences or nucleic acid sequences, but not both. Nucleic acid databases are translated on the fly by Mascot in all six reading frames.

The FASTA title line begins with a “greater than” character, followed by one or more accession strings, and an optional text string describing the entry. Apart from the use of the “greater than” character, the precise syntax of the title line varies from database to database. The title line is delimited from the sequence that follows by a platform dependent new line character.

The title line is followed by lines of contiguous sequence characters. Line lengths vary between databases; anything from 60 characters to a thousand or more. Mascot can handle lines up to 50,000 characters long. The

end of a sequence is indicated when the following line is either a new title line or the end of the file. For example:

```
.  
. .  
VY EYVRKYAEHRMLVVAEQPLHAMRKGLLDVLPKNSLEDLTAEDFRLLVNGCGEVNVQML  
ISFTSFNDESGENAEKLLQFKRWFWSIVERMSMTERQDLVYFWTSSPSLPASEEGFQMPM  
SITIRPPDDQHLPTANTCISRLYVPLYSSKQILKQKLLLAIKTKNFGFV  
>104K_THEPA (P15711) 104 KD MICRONEME-RHOPTRY ANTIGEN.  
MKFLILLFNILCLFPVLAADNHGVPQGASGVDPITFDINSNQTGPAFLTAVEMAGVKYL  
QVQHGSNVNIHRLVEGNVVIWENASTPLYTGAIVTNNDGPY MAYVEVLGDPNLQFFIKSG  
DAWVTLSEHEYLAKLQEIRQAVHIESVFSLNMAFQLENNKYEVETHAKNGANMVTFI PRN  
. . .
```

Mascot doesn't search the FASTA file directly. When a new database is recognised, Mascot Monitor uses the FASTA file to create a set of compressed files. One reason for doing this to separate the sequence string from the title line, because only the sequence string needs to be memory mapped. In the case of a database with predominantly short sequences, this greatly reduces the amount of memory required. In the case of a nucleic acid database, the limited character set allows Mascot to pack two base codes into each byte of memory. If a taxonomy filter is required, a taxonomy index is built at the same time as the file is compressed.

Sometimes, the database may not be available in FASTA format, and it will be necessary to find or write a utility program to create a FASTA file from the database native file format.

Naming Conventions and Directory Structure

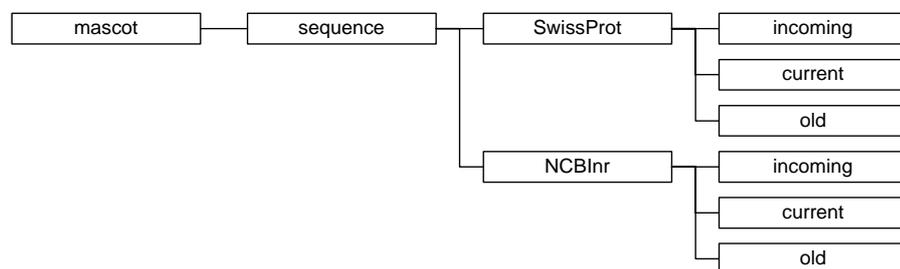
N.B. Although Microsoft Windows permits file and directory names to include spaces, file and directory names to be used by Mascot, or to appear in a URL, cannot include spaces.

By following some simple conventions in database naming, Mascot Monitor enables sequence databases to be automatically updated without any disruption to on-going searches.

The procedure followed by Monitor is that the new database is compressed and tested by running a standard search. If errors are detected in the new database, the database exchange process is abandoned. Assuming the test is successful, all new searches are performed against the new database, while searches that were in progress against the old database are allowed to continue. Once the final search against the old

database is complete, the disk file is moved into an archive directory. If the database being exchanged is memory mapped, the mapping and un-mapping are also handled automatically.

Assuming that the new database will be updated periodically, a directory structure similar to the one created for SwissProt during installation is recommended. For example:



For each database, the *incoming* directory provides a workspace for downloading and expanding a new database file. The *current* directory contains the active database, and this is where Mascot Monitor creates the memory mapped compressed files. The *old* directory is where the immediate past database file is archived ... just in case.

In the Mascot configuration, the filename for each database **must** include a wild card. This is to enable the automatic recognition and exchange of an update file. For example, the filename for the SwissProt database might be defined as `SwissProt*.fasta`. This would match to filenames that included a release number or a date stamp, e.g. `SwissProt_51.0.fasta`, or a date, e.g. `SwissProt_20070311.fasta`.

Whenever Monitor sees a file in that directory which matches to the database name and is *not* the current database, it will initiate the exchange process. This is why the wild card is important, even though you may not wish to track database dates or revision numbers.

Even if you never intend to swap a database, and have called it (say) `SwissProt.fasta`, you must still define it in the Mascot configuration using a wild card as `SwissProt*.fasta`.

Obtain a local copy of the FASTA format database

In general, you will download databases from the Internet. Download URLs and detailed configuration information for many popular databases can be found on the Matrix Science web site at http://www.matrixscience.com/help/seq_db_setup.html

To get you started, and as a service to users who do not have a fast connection to the Internet, a selection of database files is supplied on

DVD. Please note that these files will become increasingly out-of-date. If possible, you should download updates from the Internet at regular intervals. The databases on the DVD are:

1. MSDB
2. NCBIInr
3. Trembl
4. IPI_human
5. IPI_mouse

In each case, the files have been renamed to include a version or date stamp, and placed into the recommended directory structure. The procedure to enable each database is as follows:

1. Choose a suitable location for the database files
2. Unpack the files from the DVD archive
3. In some cases, unpack the additional files required to create taxonomy indexes).
4. If necessary, modify the existing Mascot configuration to identify the location of the database files.
5. Enable the new database.

Note: If you are upgrading Mascot, and use this procedure to update one or more of your databases, first stop the Mascot service (Windows) or kill `ms-monitor.exe` (Unix), and delete the old database files. Once all the new files are in place and you have updated your Mascot configuration, you will need to re-start the Mascot service / `ms-monitor.exe`

1. Choose a suitable location for the database files

The “default” location for databases is under the Mascot sequence directory, as illustrated in the directory structure, above. However, database files can be located on any local drive. If you decide to put the files in a different location, you will need to make a configuration change in step 4.

2. Unpack the files from the DVD archive

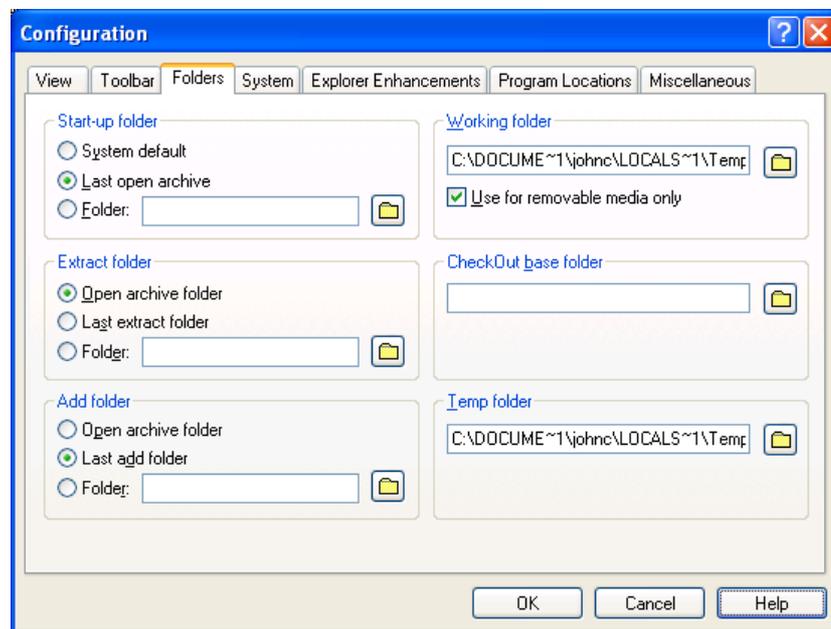
Unix: If your Mascot server is on a Unix platform, you can unpack the files using `gzip` and `tar`. If the Databases DVD is mounted as `/mnt/dvdrom` and you wish to unpack the MSDB files to `/usr/local/mascot/sequence`

```
cd /usr/local/mascot/sequence
gzip -dc /mnt/dvdrom/MSDB.tar.gz | tar xvf -
```

Windows (GUI): Many people will prefer to use a graphical utility like WinZip to unpack these archives. You will need to use WinZip 10.0 or later to unpack MSDB, because earlier versions cannot cope with files

larger than 4GB. Unfortunately, even the latest release, WinZip 11.0, cannot handle the Trembl archive, so you will have to use a command line utility, as described below.

Before trying to unpack a database, launch WinZip and choose Configuration from the Options menu. On the miscellaneous tab, clear the checkbox for 'TAR file smart CR/LF conversion'. On the Folders tab, check the location of the 'Temp folder'.

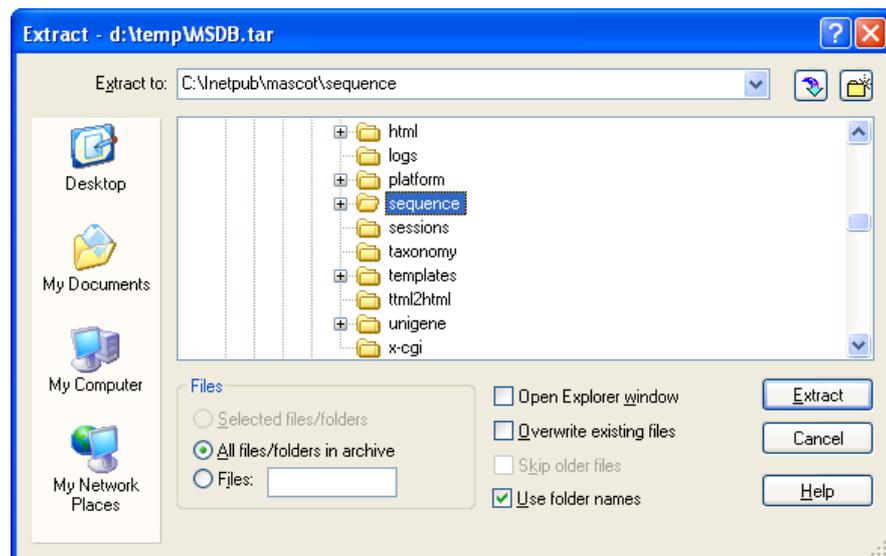


By default, this will be your temp directory, and is where WinZip will attempt to unpack the database files. Some of these are very large, (MSDB totals 6.4 GB), so you may need to change this setting to a drive with sufficient free space. Choose OK to save your changes then use WinZip to open the *MSDB.tar.gz* file. Usually, WinZip will have set your file associations so that you can open a file with a *gz* extension by double clicking it. When you open *MSDB.tar.gz*, WinZip will ask



Choose Yes, and the archive will be opened and displayed

Choose Extract, and then select your Mascot sequence directory. Make sure that 'All files' and 'Use folder names' are selected.



Windows (Command line): Download and install BsdTar from

<http://gnuwin32.sourceforge.net/packages/bsdtar.htm>

By default, the executable is installed into *C:\Program Files\GnuWin32\bin\bsdtar.exe*. If the Databases DVD is drive *D:* and you wish to unpack the Trembl files to *C:\inetpub\mascot\sequence*, open a command prompt (Windows Start menu; Programs; Accessories), and enter the following:

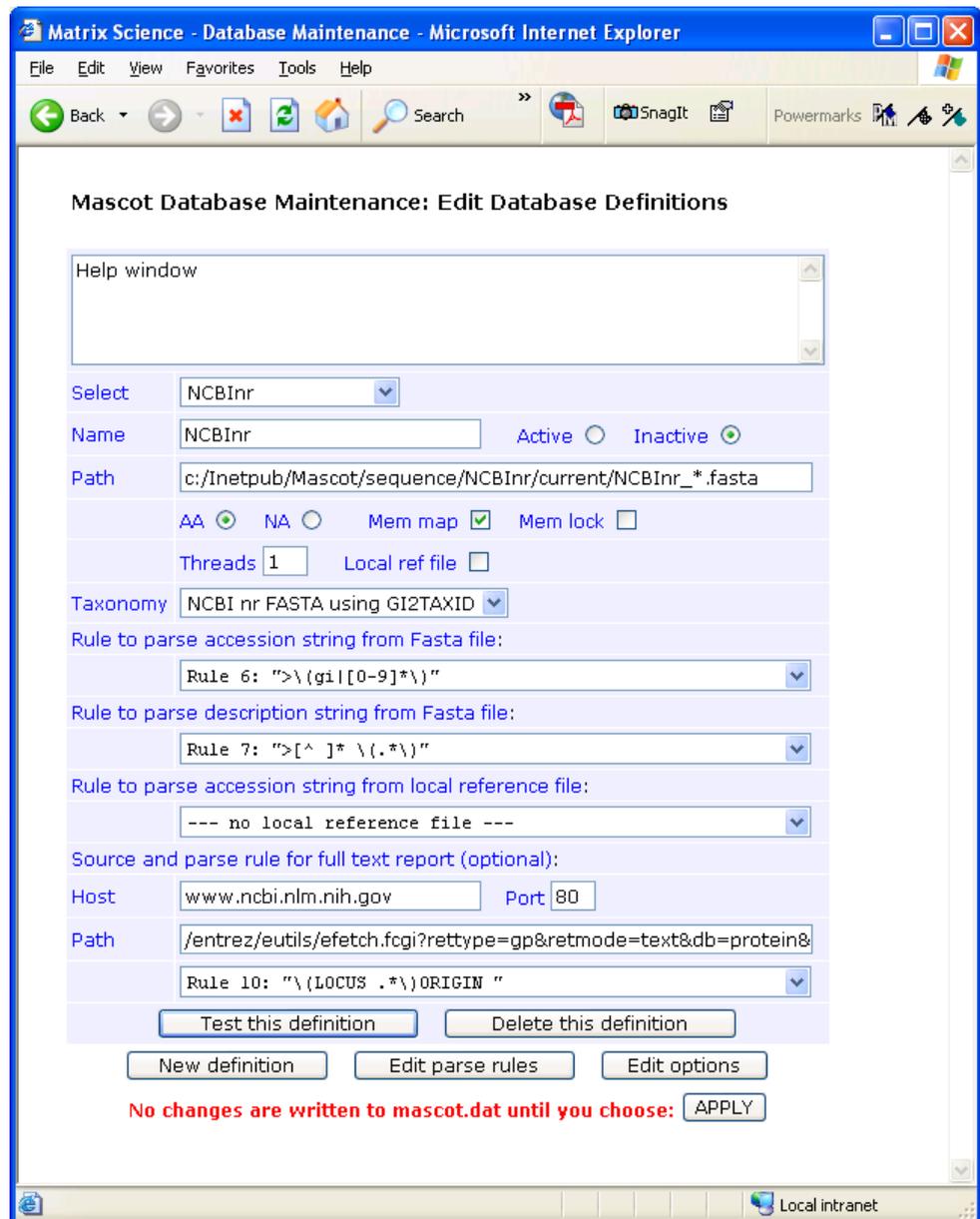
```
C:
cd \inetpub\mascot\sequence
"C:\Program Files\GnuWin32\bin\bsdtar.exe" -xvf D:\Trembl.tar.gz
```

3. Unpack the additional files

MSDB, NCBIInr, and Trembl are comprehensive adatabases containing entries from many different organisms. Mascot supports taxonomy filtering, which allows you to search a subset of the database entries, such as those for mammals or for Homo Sapiens. Building a taxonomy index requires additional files, and these are supplied on the DVD in an archive called `taxonomy.tar.gz`. Each database uses a different mix of files, but the easiest thing is to copy all the files in one go, whichever database you plan to use. Unpack the files in `taxonomy.tar.gz` using the same approach as in step 2, but copying the files into the Mascot taxonomy folder. Choose to overwrite any existing files unless you know that the files on the hard disk are more recent than those from the DVD.

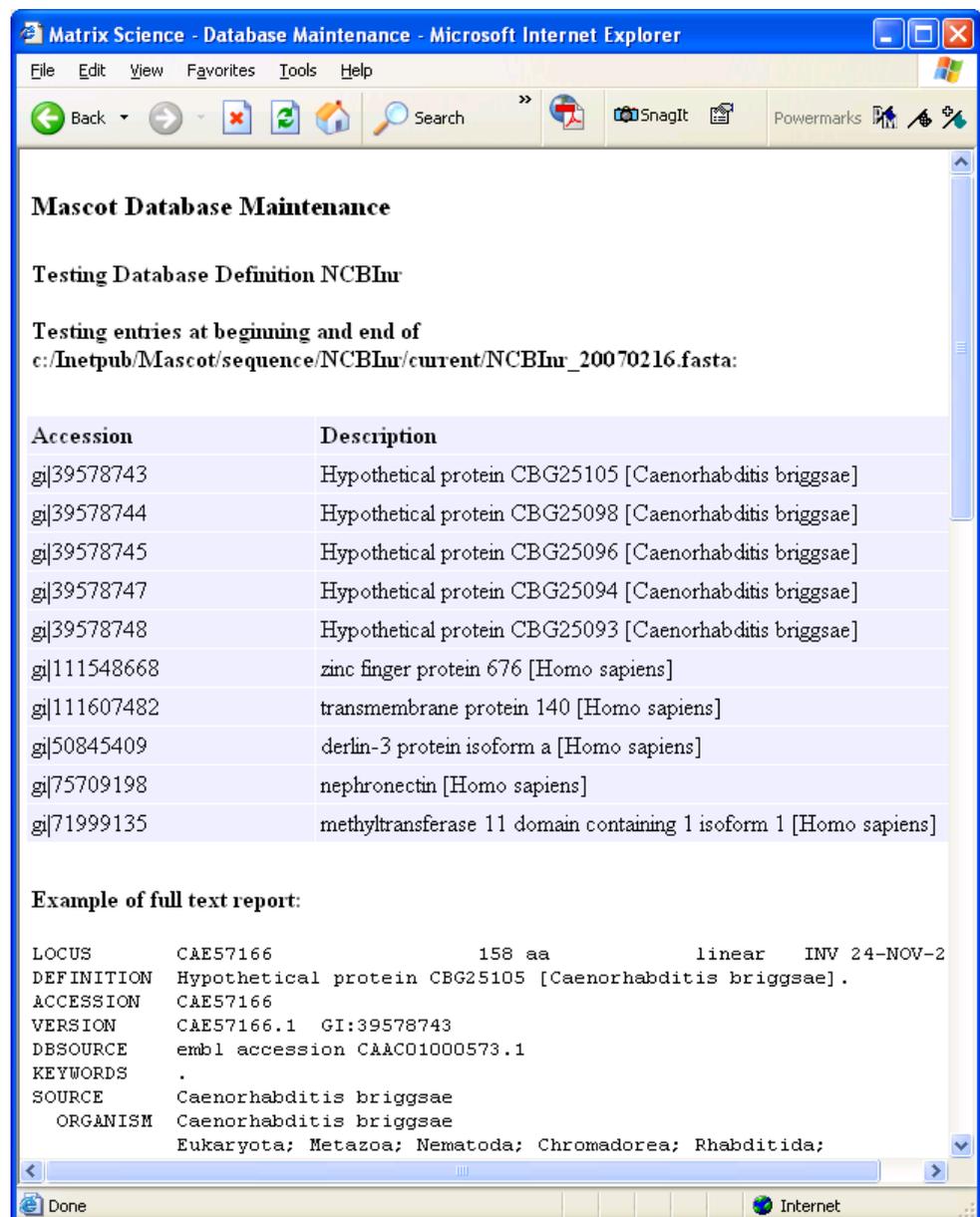
4. Modify the Mascot configuration

From your local Mascot home page, choose Configuration Editor. From the menu, choose Database Maintenance. All of these databases are pre-configured, but are not enabled. If we are installing NCBIInr, choose NCBIInr from the Select drop-down list. It will look similar to this



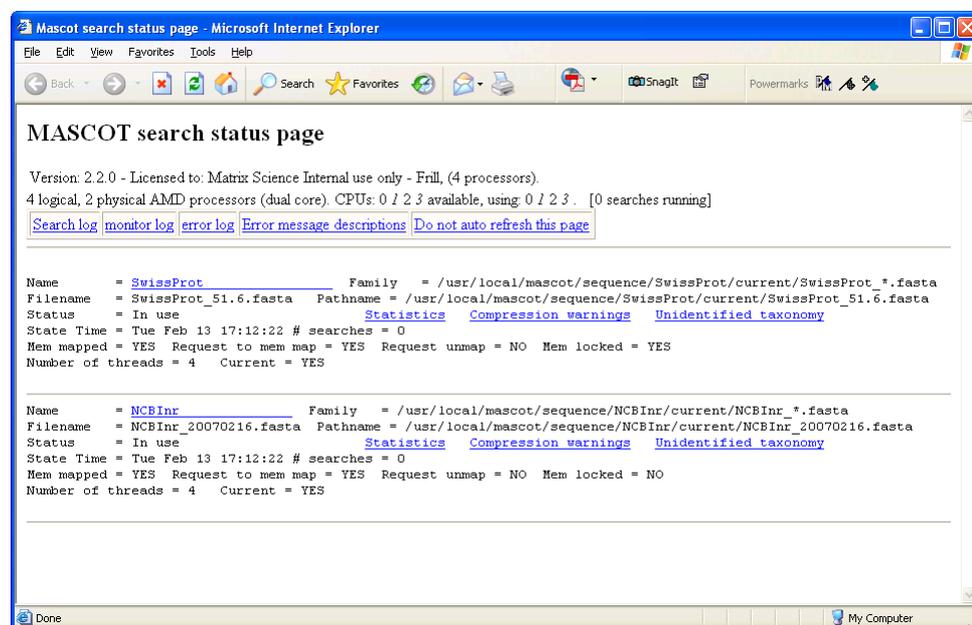
In step 2, if you copied the files to a different location, you will need to change the Path to match. Note that the slashes must be forward

slashes, even under Windows, and that there must be a wild card before the fasta file extension. Otherwise, all you need to do is change Inactive to Active. Choose Test this definition, and you should see something similar to this



If there are any error messages, these must be investigated and fixed. Choose Return to database definitions at the bottom of the page to return to the configuration form and choose Apply to save the changes.

Follow the Database Status link to monitor the database as the files are compressed and tested. The database is not available for searching until the Status is 'In Use'. (If you are upgrading Mascot, you'll have to start the Mascot service or ms-monitor.exe to view Database Status and compress the new databases.)



Troubleshooting

Proxy Server

The NCBIInr database does not have a local reference file. The default configuration is to retrieve full annotation text as required from the NCBI Entrez web site. If there is a proxy server between your Mascot server and the Internet, this may fail unless you define your proxy server in the Options section of *mascot.dat*. The relevant parameters are **proxy_server**, **proxy_username**, and **proxy_password**. Unless your proxy server uses authentication, you only need the first of these, and a typical entry in *mascot.dat* will look like this

```
proxy_server http://our-cache:3128
```

You can change this setting in Database Maintenance by choosing Edit Options at the bottom of the Database Definition form.

Permissions / Security

Mascot monitor will need to create the compressed database files in the database *current* directory, and may need to move old database files to the *old* directory. Mascot searches, running as CGI processes with very restricted privileges, need to read the files. Make sure permissions, (in Unix), or security settings, (in Windows), don't prevent this.

Files not where they are supposed to be

When you test the database, if one or the files cannot be found, this might be the result of confusion or typos. Double check that the sequence database files are exactly where the Path definition specifies. Note that the taxonomy files are shared, and go into the Mascot *taxonomy* directory, not a sequence database directory. Under Windows, remember that the directory separator must be a forward slash, not a back slash.

DVD read errors

The file sizes and checksums (as reported by cksum) are:

```
1047602745 54763210 IPI_human.tar.gz
2903524392 48053726 IPI_mouse.tar.gz
1219877599 1120629082 MSDB.tar.gz
2604784395 1112962694 NCBIInr.tar.gz
1081638331 38624927 taxonomy.tar.gz
4250782818 2103831535 Trembl.tar.gz
```

Other common problems

From the HTML help index, follow the links for *Sequence database setup* and then *Common mistakes*.

Reference

Database Maintenance

The easiest way to add a new database to Mascot is by using the Database Maintenance script. On your local Mascot home page, choose Configuration Editor then choose Database Maintenance.

Note that no changes are saved to *mascot.dat* until you explicitly choose to 'Apply' the new definitions.

Whenever changes are applied, the Database Maintenance script will automatically make backup copies of *mascot.dat*. These are numbered

sequentially from *mascot.dat.1* onwards. If you have problems after making changes, just revert to the most recent backup.

You can also edit *mascot.dat* using any text editor, but first familiarise yourself with Chapter 6, Configuration and Log Files, which contains a comprehensive description of the configuration parameters. The information in this chapter is the minimum required to set up a new database using the maintenance script.

Databases definitions form

This form can be used to modify existing database definitions or add new ones. When Mascot is installed, the only active database is SwissProt.. Other database definitions are listed on the Select list, but they are marked as inactive.

The screenshot shows a web browser window titled "Matrix Science - Database Maintenance - Microsoft Internet Explorer". The page content is "Mascot Database Maintenance: Edit Database Definitions".

The form includes the following fields and options:

- Select:** A dropdown menu currently set to "SwissProt".
- Name:** A text input field containing "SwissProt". To its right are radio buttons for "Active" (selected) and "Inactive".
- Path:** A text input field containing "C:/INETPUB/mascot/sequence/SwissProt/current/SwissProt_*.fast".
- Options:** Radio buttons for "AA" (selected) and "NA", and checkboxes for "Mem map" (checked) and "Mem lock" (unchecked).
- Threads:** A text input field containing "1".
- Local ref file:** A checkbox that is checked.
- Taxonomy:** A dropdown menu set to "Swiss-prot DAT".
- Rule to parse accession string from Fasta file:** A dropdown menu showing "Rule 31: ">[^|]*\{[^]*\}"
- Rule to parse description string from Fasta file:** A dropdown menu showing "Rule 13: ">[^]* \{.*\}"
- Rule to parse accession string from local reference file:** A dropdown menu showing "Rule 15: ""^ID \{[^]*\}"
- Source and parse rule for full text report (optional):**
 - Host:** A text input field containing "localhost".
 - Port:** A text input field containing "80".
 - Path:** A text input field containing "c:/inetpub/mascot/x-cgi/ms-getseq.exe SwissProt #ACCESSION#".
 - Rule 24:** A dropdown menu showing "\"*.*\{ID [-A-Z0-9_].*\}"

At the bottom of the form are several buttons: "Test this definition", "Delete this definition", "New definition", "Edit parse rules", "Edit options", and a large "APPLY" button. A red warning message reads: "No changes are written to mascot.dat until you choose: APPLY".

Select: Move between definitions by choosing them off this list

Name: Each database must have a unique name. The chosen name should be short and descriptive. Note that these names are case sensitive, and much confusion can be caused by creating (say) *Sprot* and *SPROT*. The name does not need to be the same as or even similar to the filename of the actual FASTA file. Allowed characters are alphanumerics and `_ . - $ % & () []`

Active / Inactive: A database that is no longer required can be marked as inactive so as to retain the definition in *mascot.dat*. If you are certain that you will never need the definition again, it can be deleted using the button at the bottom of the definition block.

Path: The location of the FASTA file is defined in the Path field. This must be the fully qualified path to the FASTA file, with a wild card in the filename for the purposes described earlier. The delimiters between directories must always be forward slashes, even if Mascot is running on a Windows system. Mascot creates its compressed files in the same directory as the original FASTA file.

AA / NA: A pair of radio buttons is used to specify whether the database is amino acid (protein) or nucleic acid.

Mem map: Database files should always be memory mapped. Unlike memory locking, this does not consume physical RAM.

Mem lock: Memory mapped files can be locked in memory, but only if the computer has sufficient RAM. Having a database locked in memory means that it can never be swapped out to disk, ensuring maximum possible search speed. Of course, there must also be sufficient RAM for the operating system, (Windows consumes approximately 60 Mb), anything from tens to hundreds of Mb for each Mascot search, and space for any other applications which might be running.

If you try to lock databases into RAM when there isn't room, this will not be a major problem. The locking will fail, generate an error message, and Mascot will carry on regardless. A more serious problem is when there is just sufficient RAM to lock the databases, but none left over for searches or other applications. In this case, the whole system will slow down and the hard disk will be observed to be "thrashing". Eventually, the system is likely to hang or crash.

The best policy is to memory lock only smaller databases that are frequently used.

Threads: A Mascot search can use multiple threads. If you are running in cluster mode, 'Threads' must be set to 1. Otherwise, if you have single core processors with no hyperthreading, specify the same number of threads as processors in your Mascot licence. If your processors have

hyperthreading, double the number and if you have dual-core processors, double it again.

Local ref file: The FASTA database file must be available locally. For certain databases (MSDB, OWL, SwissProt, Trembl, etc.), it is also possible to have a local reference file, from which full text information can be taken for a 'Protein View' report. If you intend to have a local full text file, check the 'Local ref file' box. Otherwise, clear it.

Taxonomy source: Some databases have pre-defined taxonomy rules. If you are adding a new database, and taxonomy rules are not defined, refer to Chapter 9 for further information. Some databases, like SwissProt, have a choice of taxonomy rules according to whether you have a local full text file or not.

If you don't want to use taxonomy, select '— None —'. This will speed up the time taken to bring the database on-line.

Parse Rules: Apart from starting with a 'greater than' symbol, the precise syntax of the FASTA title line varies from database to database. For this reason, Mascot uses Basic Regular Expressions to define how the accession string and the short description should be parsed from the FASTA title line. The rules for using Basic Regular Expressions are described in detail in Appendix A and some additional information is provided later in this chapter.

Two or three fields are used to define the regular expressions used to parse information from the database files. The first rule defines how to parse an accession string from the FASTA file title line. The second defines how to parse a short description string from the FASTA file title line. The third defines how to parse an accession string from a local reference file. If there is no local reference file, the third rule should be set to '— No local reference file —'.

If you are adding a new database, some experimentation may be required to choose the correct rules. In most cases, you will find that a suitable regular expression is already defined; it is just a case of choosing it. The procedure is described below under 'Testing a database definition'.

Source and parse rules for full text report

This section defines how Mascot can retrieve a full text report (references, feature table, etc.) for use in 'Protein View'. If full text is not required or not available, just leave all fields blank.

Host: The report can come from a local reference file, or from a remote resource such as the NCBI web server. In the case of a local file, such as the SwissProt .DAT file, specify localhost. Otherwise, specify a remote host (webserver) from which a report can be obtained by HTTP.

Port: Always specify 80 unless a remote host (webserver) is known to operate on a non-standard port.

Path: For a remote host, this path completes the URL required to retrieve a report. For localhost, specify the path to ms-getseq.exe followed by the arguments required to retrieve a report. #ACCESSION# is a placeholder for an accession string. Folders must be delimited by forward slashes, even on a Windows system.

Parse rule: This parse rule is used to select the text that will appear in 'Protein View'. If the host is a remote host, you'll probably want to remove the HTML tags.

Testing the definition

Whenever creating or modifying a database definition, it is very important to 'Test this definition'. This runs a number of tests on the consistency and syntax of the various settings. The script will also read the first five and last five entries from the FASTA file (and local reference file if defined) and test the parse rules. A test report for SwissProt will look similar to this:

Mascot Database Maintenance

Testing Database Definition Sprout

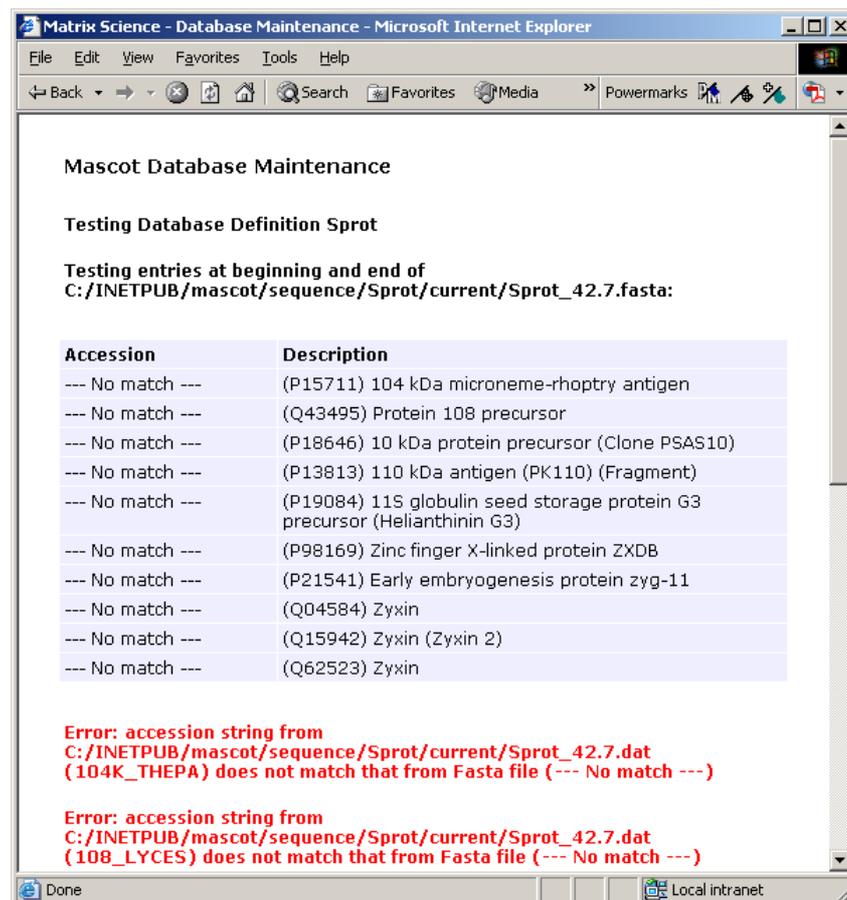
**Testing entries at beginning and end of
C:/INETPUB/mascot/sequence/Sprot/current/Sprot_42.7.fasta:**

Accession	Description
104K_THEPA	(P15711) 104 kDa microneme-rhoptry antigen
108_LYCES	(Q43495) Protein 108 precursor
10KD_VIGUN	(P18646) 10 kDa protein precursor (Clone PSAS10)
110K_PLAKN	(P13813) 110 kDa antigen (PK110) (Fragment)
11S3_HELAN	(P19084) 11S globulin seed storage protein G3 precursor (Helianthinin G3)
ZXDB_HUMAN	(P98169) Zinc finger X-linked protein ZXDB
ZY11_CAEEL	(P21541) Early embryogenesis protein zyg-11
ZYX_CHICK	(Q04584) Zyxin
ZYX_HUMAN	(Q15942) Zyxin (Zyxin 2)
ZYX_MOUSE	(Q62523) Zyxin

**Local reference file tests OK
(C:/INETPUB/mascot/sequence/Sprot/current/Sprot_42.7.dat)**

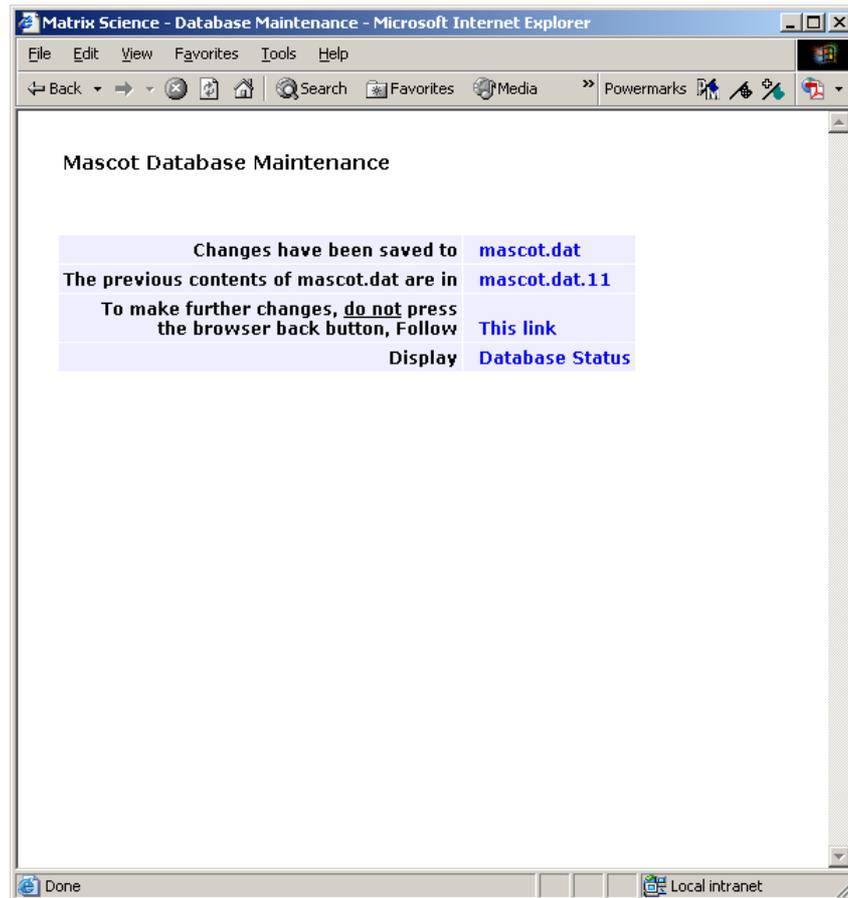
Return to database definitions

If any problems are encountered, there will either be explicit error messages, or the failure will be apparent from the test results:



Saving the changes

Use the button at the bottom of the test report to return to the database definitions form and choose 'Apply'. All being well, the following page will be displayed:



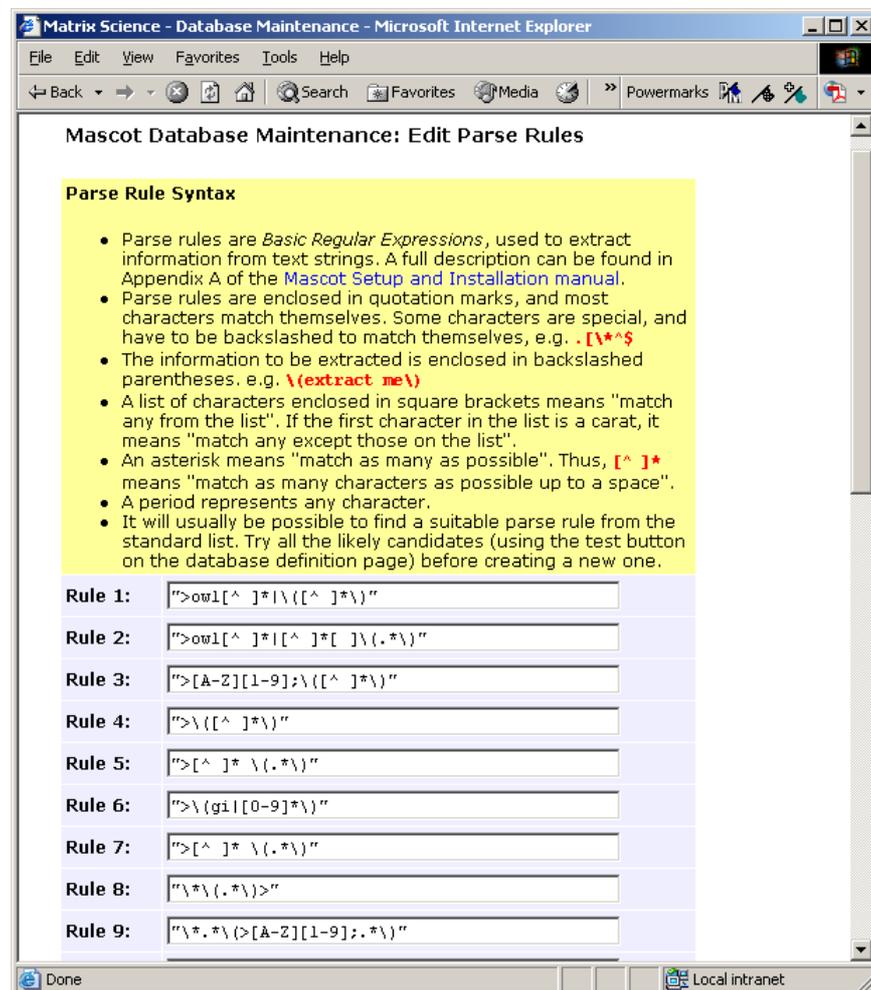
You can then follow the link to the Mascot database status page and monitor progress as the new database is brought on-line.

If all is well, the status line for the new database will display the following messages:

```
Creating compressed files
Running 1st test
First test just run OK
Trying to memory map files
Just enabled memory mapping
In Use
```

Parse Rules

New parse rules can be added, and existing ones modified or deleted by clicking on the 'Edit parse rules' button.



Several major databases have adopted the NCBI syntax for accession strings (or sequence identifiers). This is described in the following document:

<ftp://ftp.ncbi.nih.gov/blast/db/blastdb.html>

For a database that includes multiple accession strings, such as NCBIInr, the simplest approach is to write a rule that takes either the first or last accession string. Mascot only requires that the accession string is unique within the database.

There are help pages for many of the most popular public databases, which describe suitable parse rules. These can be found on the Matrix Science web site at http://www.matrixscience.com/help/seq_db_setup.html. If you are adding a new database, some experimentation may be required to choose the correct rules. In most cases, you will find that a suitable regular expression is already defined; it is just a case of choosing it.

If the Fasta title line syntax is unusual, you'll need to examine several entries to get a sense of what might work. If the database file is a large one, it may not be a good idea to open it in a standard word processor or text editor. Most platforms support a command line utility called `more` that can be used to browse a file of any size.

A simple rule that takes everything between the ">" symbol and the first space as the accession will often do the trick. Everything after the first space can be treated as the description. These rules are pre-defined in `mascot.dat` as rules 4 and 5.

```
">\ ([^ ]*\ )"
">[^ ]* \ (.*\)"
```

Database updating

When a new release of a database becomes available, it should be copied or downloaded into the *incoming* directory. In many cases, the downloaded file will have to be de-compressed. The filename may or may not be constant from release to release.

The FASTA database should then be copied into the current directory using a name that *does not* match to the database name. The reason for this is that it can take several seconds to copy a large file, and you don't want to trigger off the exchange process until the entire file has been copied.

Finally, when you are ready to swap databases, rename the database to its final name, and Mascot Monitor will begin the exchange process. Progress can be monitored from a web browser using the Mascot Database Status page

If you are using a local reference file, copy and rename this file first. Otherwise, the exchange process will be triggered off by the appearance of the FASTA file, but will immediately fail because the new reference file is not yet available.

Database update utility: db_update.pl

Overview

The purpose of this script is to download database updates for Mascot, so that the whole process can be automated using Unix *Cron* or Windows *Scheduled Tasks*. There is some complexity to achieving this, and functionality includes:

- downloading a fasta file plus optional associated files such as a full text reference file or release notes
- optionally downloading one or more taxonomy indexes
- handling variable filenames via wild cards
- uncompressing, unpacking, renaming and moving the files
- time or version stamping
- downloading a file only if a new one is available; resuming an interrupted download
- passive FTP through a firewall; HTTP proxy server authentication
- special processing, such as splice variant expansion of Swiss-prot using varsplic

Usage

```
db_update.pl DB
```

Where DB is keyword predefined in the script header. For example

```
db_update.pl NCBIInr_from_NCBI
```

Installation

db_update.pl should be placed in the Mascot bin directory. The following utilities are required:

wget <http://www.gnu.org/software/wget/wget.html>

tar <http://www.gnu.org/software/tar/tar.html>

gzip <http://www.gnu.org/software/gzip/gzip.html>

tar and gzip are likely to be present on any Unix system. Windows binaries for all three utilities can be downloaded from <http://gnuwin32.sourceforge.net/packages.html>

All three utilities should be installed into a directory on the system search path, so they can be executed from any directory, without having to provide path information.

Configuration

General

Open `db_update.pl` in a text editor and read through the header. You may need to modify the path to the Mascot root directory. Note that several definitions are specified independently for Unix and Windows, to minimise the need for editing. You only need to change the definitions for the platform you are using.

If you were unable to install the utilities into a directory on the system search path, you will need to add full path information to the relevant definition. For example:

```
$gzip = "$MASCOT/bin/gzip.exe"; ✓ correct
```

```
$gzip = "/usr/home/peter/wget"; ✓ correct
```

```
$gzip = "./wget.exe"; ✗ wrong (relative path)
```

Note that Windows paths must be specified using forward slashes, and cannot contain embedded spaces. You can specify a drive letter, but not a UNC path. For example:

```
$gzip = "C:/peter/download/gnu/wget.exe"; ✓ correct
```

```
$gzip = "D:\private\wget.exe"; ✗ wrong (back slashes)
```

```
$gzip = "C:/Program Files/gzip.exe"; ✗ wrong (spaces)
```

```
$gzip = "\\myserver\c$\utils\gzip.exe"; ✗ wrong (UNC)
```

If there is an HTTP proxy server between the Mascot server and the internet, and this proxy server requires authentication, you should uncomment and modify the `$wget_options` definitions for `-proxy-user` and `-proxy-passwd`. On Unix systems, the proxy server URL will normally be found from the environment. On Windows systems, you may need to specify it with additional `$wget_options`.

Database Update Definition Blocks

Several common database update definition blocks are pre-configured, and you may not need to add or change anything before using the script. A particular definition block is chosen by means of a keyword argument. Execute the script without any arguments to get a list of recognised keywords:

```
MSDB_from_NCBI
MSDB_from_EBI
NCBINr_from_NCBI
EST_human_from_NCBI
EST_mouse_from_NCBI
EST_others_from_NCBI
SwissProt_fasta_only_from_EBI
SwissProt_complete_from_EBI
SwissProt_varsplic_from_EBI
Trembl_complete_from_EBI
IPI_human_from_EBI
IPI_mouse_from_EBI
UniRef100_Fasta_from_EBI
```

Take NCBINr_from_NCBI as an example:

```
#
# NCBINr_from_NCBI
#
} elsif ($ARGV[0] eq "NCBINr_from_NCBI") {
  $db_name = "NCBINr";
  $local_incoming_directory = "$SEQUENCE/NCBINr/incoming";
  $local_current_directory = "$SEQUENCE/NCBINr/current";
  $fasta_file_url = "ftp://ftp.ncbi.nih.gov/blast/db/
  FASTA/nr.gz";
  $taxonomy_file_url[0] = "ftp://ftp.ncbi.nih.gov/pub/taxonomy/
  taxdump.tar.gz";
  $taxonomy_file_url[1] = "ftp://ftp.ncbi.nih.gov/pub/taxonomy/
  gi_taxid_prot.dmp.gz";
```

The keyword NCBINr_from_NCBI is descriptive of both the database and the download source. \$db_name is the Mascot name for the database (n.b. case sensitive). \$local_incoming_directory is the directory where the database files will be downloaded and processed. When processing is complete, the new database files will be moved to \$local_current_directory, which will trigger Mascot to automatically swap to the new database. In this particular case, there is only a Fasta database file.

Since there is no explicit version information, the new database will be time-stamped with the file modification date in ISO format. For example, NCBINr_20020625.fasta.

There are two sets of taxonomy files that need to be updated at the same time as the database. The first taxonomy file URL is always \$taxonomy_file_url[0], and any additional files are numbered sequentially.

If the HTTP or FTP server requires authentication, this can be encoded in the URL

```
"ftp://user:password@host/path"
```

```
"http://user:password@host/path"
```

Now, a slightly more complex example:

```
#
# SwissProt_complete_from_EBI
#
} elsif ($ARGV[0] eq "SwissProt_complete_from_EBI") {
  $db_name = "SwissProt";
  $local_incoming_directory = "$SEQUENCE/SwissProt/incoming";
  $local_current_directory = "$SEQUENCE/SwissProt/current";
  $fasta_file_url = "ftp://ftp.ebi.ac.uk/pub/databases/
    uniprot/current_release/knowledgebase/complete/
    uniprot_sprot.fasta.gz";
  $name_file_url = "ftp://ftp.ebi.ac.uk/pub/databases/
    uniprot/current_release/knowledgebase/complete/reldate.txt";
  $reference_file_url = "ftp://ftp.ebi.ac.uk/pub/databases/
    uniprot/current_release/knowledgebase/complete/
    uniprot_sprot.dat.gz";
  $taxonomy_file_url[0] = "ftp://ftp.ncbi.nih.gov/pub/taxonomy/
    taxdump.tar.gz";
  $taxonomy_file_url[1] = "ftp://ftp.ebi.ac.uk/pub/databases/
    uniprot/current_release/knowledgebase/complete/docs/
    speclist.txt";
  $version_regex = 'Swiss-Prot\s+.*Release\s+([\d\.]+)';
}
```

In addition to a Fasta file, Swiss-Prot has a reference file containing annotation text (`uniprot_sprot.dat`) and release notes containing version information (`reldate.txt`). The string `$version_regex` is an extended regular expression describing how to parse the version number from the release notes file. Note that this string must be enclosed in single apostrophes, not double. For details of how to construct extended regular expressions, refer to your Perl documentation.

If the regular expression succeeds, the version number will be used to identify the database. If it fails, an ISO time-stamp will be used instead.

If the name or location of a download file changes, you will need to update the corresponding definition block. If you want to add a new database, the easiest way is to make a copy of a similar looking definition block and then modify it. Note that each definition block must start with an `elsif` statement, as shown above, and the whole block must be placed before the final `else` statement.

You can download files from HTTP servers, but this should only be done if no FTP server is available. Downloads from HTTP servers do not allow for resumption of failed downloads, do not allow wild cards to be used in the filename, and will always proceed, even if the file has been downloaded on a previous occasion.

UniGene

The database definition blocks for EST_human, EST_mouse, and EST_others include paths for downloading a variety of NCBI UniGene indexes. These indexes can be used by the master results report to cluster EST matches into gene families. To enable downloading of UniGene indexes, simply uncomment the definition line for `$local_unigene_directory`.

Testing a Database Definition

Before adding a new `db_update.pl` entry to Unix *Cron* or Windows *Scheduled Tasks*, it is essential to test it.

Unix

To properly test the functionality, you should execute the script at a shell prompt when logged on as the proposed owner of the Cron job, and from a directory other than the one in which the script is located. This will ensure that directory permissions are correct and the paths can be resolved. For security, the owner of the Cron job should not be root.

Windows

To properly test the functionality, you should execute the script from a command prompt and from a directory other than the one in which the script is located. This will ensure that directory permissions are correct and the paths can be resolved.

Automation

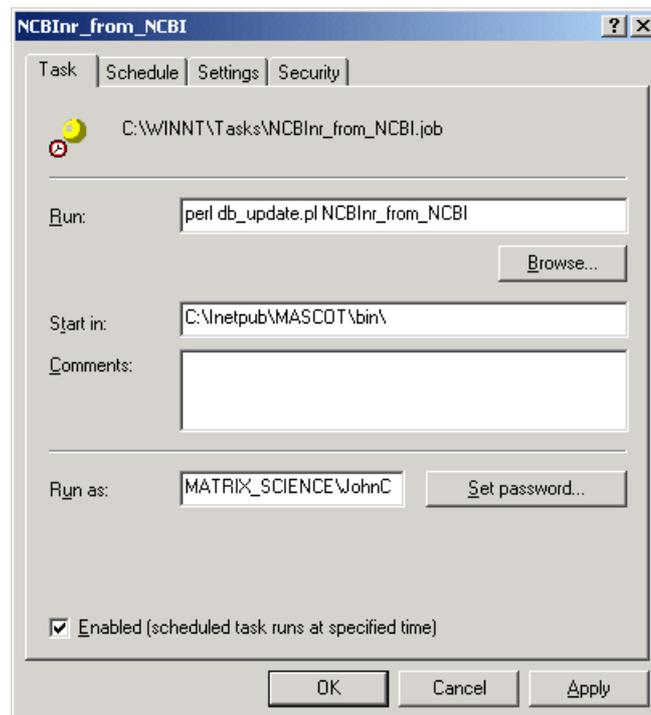
Unix

Once the script has been found to function correctly for a particular definition block, an entry can be added to Cron. As a rule, you should stagger database updates through Mascot server quiet periods. Trying to update all the databases simultaneously will prolong the download times and may slow down any Mascot searches currently in progress.

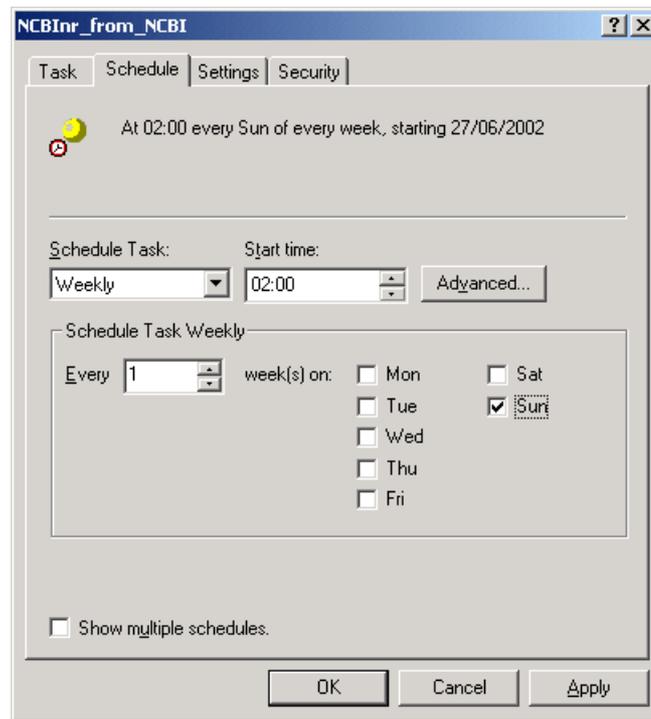
Windows 2000 / XP / 2003

Windows *Scheduled Tasks* provides a mechanism for executing `db_update.pl` at a predetermined time. You can locate the Scheduled Tasks folder from the Start menu; Programs; Accessories; System Tools. There is a wizard to add new tasks, but it is easier to add a new task from the file menu then edit the properties.

It's a good idea to use the database keyword for the task name. Right click the new task, and choose properties. In the *Run* field, enter perl followed by the script name followed by the database keyword. Set the *Start in* directory to the location of the script:



The schedule can be whatever you wish. For example:



Press OK to save the task. You will be asked for the password of the user who owns the task. The same process can be repeated to update other databases. As a rule, you should stagger database updates during Mascot server quiet periods. Trying to update all the databases simultaneously will prolong the download times and may slow down any Mascot searches currently in progress.

The format of the Cron entries is identical to a Unix crontab file, consisting of lines of six fields. The fields are separated by spaces or tabs. The first five are integer patterns that specify the following: minute (0-59), hour (0-23), day of the month (1-31), month of the year (1-12), day of the week (0-6 with 0=Sunday).

A pattern is either an integer or two integers separated by a minus sign (meaning an inclusive range). Note that days may be specified in two different ways (day of the month and day of the week). If both are specified as a list of elements, both are adhered to. Up to two fields can contain asterisks, meaning all legal values.

The sixth field is a string that is executed at the specified times. The string must be on a single line, and the entire string, up to the end of the line, is passed to the system for execution. The part of the string up to the first space must be an executable. The remainder of the line will be passed to the executable as parameters. Note the use of forward slashes.

Miscellaneous

A single log file is maintained for all instances of `db_update.pl`. The location is defined by `$local_log_file` in the script header.

Each file downloaded by FTP is listed in a file called `.history`, located in the corresponding incoming directory. This is used to prevent a given file being downloaded more than once. If you want to defeat this mechanism, simply delete or edit the `.history` file.

Special Processing

varsplic

`varsplic` is an EBI utility that generates additional sequences representing the splice variants, variants and conflicts described in Swiss-Prot annotations.

`ftp://ftp.ebi.ac.uk/pub/software/swissprot/varsplic/`

`ftp://ftp.ebi.ac.uk/pub/software/swissprot/Swissknife/`

The standard `varsplic.fas` file available for download from ExPasy corresponds to:

```
varsplic.pl -input sprot.dat -fasta varsplic.fas -dbcode
```

This contains just splice variants, and doesn't include the original Swiss-Prot entries. To produce a database suitable for Mascot searches, complete with an equivalent dat file, `db_update.pl` executes `varsplic.pl` with the following parameters:

```
varsplic.pl -input sprot.dat -pseudo varsplic.dat -  
fasta varsplic.fas -which full -uniqids 1 -count -  
varsplic -variant -conflict
```

This produces a database with the original entries plus all splice forms, variants and conflicts expanded. The title line for an original entry looks like

```
>P13813 110 kDa antigen (PK110) (Fragment)
```

The title line for the original 'parent' entry, from which new entries have been generated, looks like:

```
>P80438 (12KD_MYCSM) 12 kDa protein (Fragment) 12 kDa protein  
(Fragment)
```

The title line for a new, expanded entry looks like:

```
>P15455-00-00-00 (12S1_ARATH) Splice isoform Displayed; Variant  
Displayed; Conflict Displayed; from P15455 12S seed storage pro-  
tein precursor
```

To use `varsplic`, the location of the perl script must be defined in the `db_update.pl` header. The `Swissknife` package is also required, and must be placed on the system `$PERLLIB` path. (Generally `C:\Perl\site\lib` on a Windows system).

6

Configuration & Log Files

Configuration Files

Mascot configuration files are located in the *mascot/config* directory:

unimod.xml defines mass values and modifications, including substitutions

enzymes defines enzyme cleavage specificity

fragmentation_rules specifies which fragment ion series correspond to defined instrument types

mascot.dat contains general configuration information, such as the paths to the sequence databases.

taxonomy specifies the taxonomy filter choices for the search form (described in Chapter 9)

quantitation.xml defines quantitation methods

odelist.txt configures the systems belonging to a Mascot cluster (described in Chapter 11)

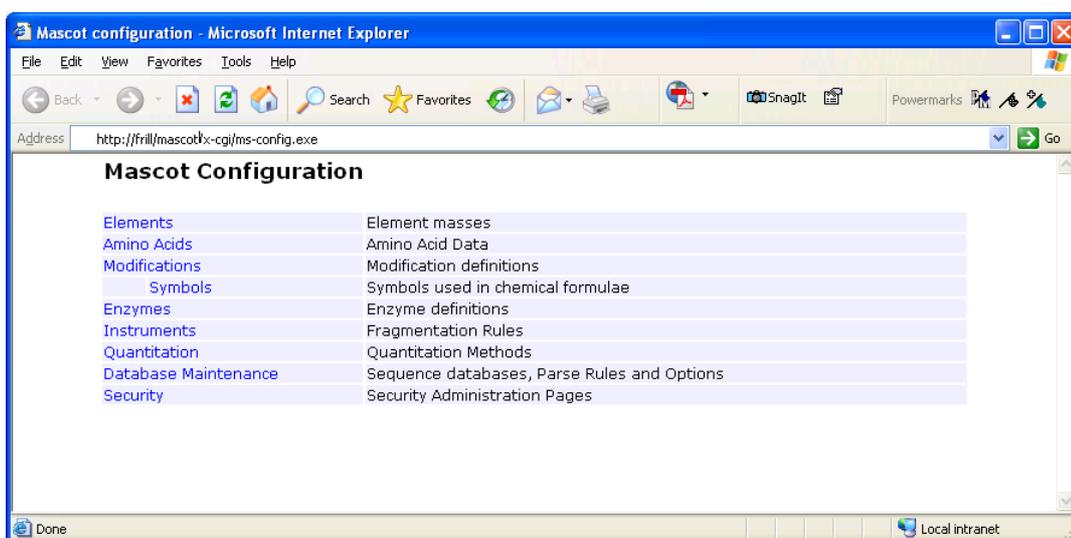
user.xml, *group.xml*, *security_options.xml*, and *security_tasks.xml* are the configuration files for Mascot security, described in Chapter 12

mod_file, *masses*, and *substitutions* are obsolete configuration files that are created on the fly from *unimod.xml* to support third party applications that expect to find these files.

A browser-based Configuration Editor is provided to view and edit these files. These files are all text files, so can also be edited in any text editor. If you choose to edit the files, exercise care and always make a backup first, because seemingly small errors can render Mascot unusable.

Configuration Editor

The local Mascot home page contains a link to the Configuration Editor. The top-level page is a menu.



unimod.xml

The first four menu items: Elements, Amino acids, Modifications, and Symbols, are interfaces to different aspects of *unimod.xml*. It should only be necessary to make changes to *unimod.xml* in exceptional circumstances. An example might be if you wanted to add a modification that was confidential. Otherwise, if you wish to add a new modification and it is not confidential, better to add it to the public Unimod database, www.unimod.org, and later download an updated configuration file from the Unimod help page. By going this route, you share the new modification with others, and benefit in turn from other people's updates.

Most of the pages of the Configuration Editor are self-explanatory. Where necessary, help text is displayed when the mouse rolls over a hyperlink.

Mascot Configuration: Modifications

Title	Monoisotopic	Average	Composition			
PEO-Iodoacetyl-LC-Biotin	414.193691	414.5196	H(30) C(18) N(4) O(5) S	Copy	Delete	Print
PET	121.035005	121.2028	H(7) C(7) N O(-1) S	Copy	Delete	Print
PGA1-biotin	660.428442	660.9504	H(60) C(36) N(4) O(5) S	Copy	Delete	Print
Phe->Cys	-44.059229	-44.0310	H(-4) C(-6) S	Copy	Delete	Print
Phe->Ile	-33.984350	-34.0162	H(2) C(-3)	Copy	Delete	Print
Phe->Ser	-60.036386	-60.0966	H(-4) C(-6) O	Copy	Delete	Print
Phe->Tyr	15.994915	15.9994	O	Copy	Delete	Print
Phe->Val	-48.000000	-48.0428	C(-4)	Copy	Delete	Print
Phenylisocyanate	119.037114	119.1207	H(5) C(7) N O	Copy	Delete	Print
Phenylisocyanate:2H(5)	124.068498	124.1515	2H(5) C(7) N O	Copy	Delete	Print
Phospho	79.966331	79.9799	H O(3) P	Copy	Delete	Print
Phosphoadenosine	329.052520	329.2059	H(12) C(10) N(5) O(6) P	Copy	Delete	Print
Phosphoguanosine	345.047435	345.2053	H(12) C(10) N(5) O(7) P	Copy	Delete	Print
PhosphoHex	242.019154	242.1205	H O(3) P Hex	Copy	Delete	Print
PhosphoHexNAc	283.045704	283.1724	H O(3) P HexNAc	Copy	Delete	Print
Phosphopantetheine	340.085794	340.3330	H(21) C(11) N(2) O(6) P S	Copy	Delete	Print
PhosphoribosyldephosphoCoA	881.146904	881.6335	H(42) C(26) N(7) O(19) P(3) S	Copy	Delete	Print
PhosphoUridine	306.025302	306.1660	H(11) C(9) N(2) O(8) P	Copy	Delete	Print
Phycocyanobilin	586.279135	586.6780	H(38) C(33) N(4) O(6)	Copy	Delete	Print
Phycocerythrobilin	588.294785	588.6939	H(40) C(33) N(4) O(6)	Copy	Delete	Print

Page 20 of 25 Go to page << >> Page size

Biotinyl-iodoacetamidyl-3,6-dioxaoctanediamine
Alternative name 1: Pierce EZ-Link PEO-Iodoacetyl Biotin

For modifications, more detailed help can be found in the Unimod help pages. This is also the place to find details of the file format, which is fully defined by a schema called unimod_1.xsd.

Enzymes

Mascot Configuration: Enzymes

Title	Sense	Cleave at	Restrict	Independent	Semispecific	
Trypsin	C-Term	KR	P	no	no	Edit Delete
Arg-C	C-Term	R	P	no	no	Edit Delete
Asp-N	N-Term	BD		no	no	Edit Delete
Asp-N_ambic	N-Term	DE		no	no	Edit Delete
Chymotrypsin	C-Term	FLWY	P	no	no	Edit Delete
CNBr	C-Term	M		no	no	Edit Delete
CNBr+Trypsin	C-Term	M		no	no	Edit Delete
	C-Term	KR	P	no	no	Edit Delete
Formic_acid	C-Term	D		no	no	Edit Delete
Lys-C	C-Term	K	P	no	no	Edit Delete
Lys-C/P	C-Term	K		no	no	Edit Delete
PepsinA	C-Term	FL		no	no	Edit Delete
Tryp-CNBr	C-Term	KMR	P	no	no	Edit Delete
TrypChymo	C-Term	FKLRWY	P	no	no	Edit Delete
Trypsin/P	C-Term	KR		no	no	Edit Delete
V8-DE	C-Term	BDEZ	P	no	no	Edit Delete
V8-E	C-Term	EZ	P	no	no	Edit Delete
semiTrypsin	C-Term	KR	P	no	yes	Edit Delete
LysC+AspN	N-Term	BD		no	no	Edit Delete
	C-Term	K	P	no	no	Edit Delete
None						

Enzyme 'None' is a special case, which cannot be modified or deleted. All the other enzyme definitions can be edited or deleted, and new ones added.

The edit page allows you to test a new enzyme definition against a protein

Edit Enzyme: CNBr+Trypsin

General

Title: CNBr+Trypsin

Independent:

Semispecific:

Components

#	Sense	Cleave At	Restrict	Delete
1	C-Term	M		<input type="checkbox"/>
2	C-Term	KR	P	<input type="checkbox"/>

Test

Protein :
 MSEELSQKPSSAQSLSLREGNRNRPFLSLSQREGFRFFPSLSLSEDRGRKFSFLSMFSFLM
 PLLEVIKIIIASVASVIFVGFACVTLA GSAAALVSTP VFIIIFSPVL VPATIA TVLATGFTAG
 GSFGATALGLIMWLVKRRMGVKPKDNPPAPAGLPPNSGAGAGGAQSLIKKSKAKSGGLK

#	Start	End	Peptide
1	1	1	M
2	2	18	SEELSQKPSS AQSLSLR
3	19	21	EGR
4	22	23	NR
5	24	32	FPFLSLSQR
6	33	35	EGR
7	36	45	FFPSLSLSEK
8	46	48	DGR
9	49	49	K
10	50	55	FSFLSM
11	56	60	FSFLM
12	61	67	PLLEVIK
13	68	136	IIIASVASVI FVGFACVTLA GSAAALVSTP VFIIIFSPVL VPATIA TVLATGFTAGSF GATALGLIM
14	137	140	WLVK

File format (enzymes)

Each cleavage agent is defined by a block of lines. Blocks are delimited from one another by a line containing an asterisk. Each line in a block starts with a keyword.

```
Title:Trypsin
Cleavage:KR
Restrict:P
Cterm
*
Title:Asp-N
Cleavage:DB
Nterm
*
```

The first line of each block must start with the `Title:` keyword, followed by a text string that is used to identify the cleavage agent in forms and reports. The definition should be short and self-explanatory. It should only include alphanumeric characters and spaces. Internal spaces are significant.

Each block must also include a line starting with the keyword `Cleavage:` followed by a list of the residues that identify the cleavage site.

Optionally, a block can include a line starting with the keyword `Restrict:` followed by a list of the residues which prevent cleavage if present adjacent to the potential cleavage site.

Finally, the block must include either the keyword `Cterm` or `Nterm` to define whether cleavage occurs on the C terminal or N terminal side of the specified residues.

This syntax can be extended to support multiple cleavage specificities, enabling enzyme mixtures to be modelled, or mixed C-term and N-term cutters. This is achieved by appending zero-based index numbers in square brackets to the keywords `Cleavage`, `Restrict`, `Cterm`, and `Nterm`. For example:

```
Title:CNBr+Trypsin
Cleavage[0]:M
Cterm[0]
Cleavage[1]:KR
Restrict[1]:P
Cterm[1]
Independent:0
*
```

The use of index numbers is optional when only one specificity is defined, but required when there are multiple specificities, as in this example.

For a definition with multiple specificities, if the keyword `Independent` appears and is given a value of 1, this means that the specificities should be treated as if independent digests had been performed on separate sample aliquots and the resulting peptide mixtures combined. Thus, any given peptide will conform to the specificity of one cleavage type only. In the case of `CNBr+Trypsin`, if `Independent` was set to 1, you would not find any peptides resulting from cleavage after K or R at one end, and

cleavage after M at the other. When `Independent` is omitted or given a value of 0, the specificities are combined, as if the reagents had been applied simultaneously or serially to a single sample aliquot. The keyword `Independent` does not take an index.

```
Title:semiTrypsin
Cleavage[0]:KR
Restrict[0]:P
Cterm[0]
SemiSpecific:1
*
```

If the keyword `SemiSpecific` appears and is given a value of 1, this means that any given peptide need only conform to the cleavage specificity at one end. The other end can result from non-specific cleavage. When `SemiSpecific` is omitted or given a value of 0, peptides are required to conform to the cleavage specificity at both ends. The keyword `SemiSpecific` does not take an index.

Instruments

Ion series	Default	ESI QUAD TOF	MALDI TOF PSD	ESI TRAP	ESI QUAD	ESI FTICR	MALDI TOF TOF	ESI 4SECTOR	FTMS ECD	ETD TRAP	MALDI QUAD TOF
1+	X	X		X	X	X		X	X	X	X
2+	X	X		X	X	X		X	X	X	X
2+ (precursor>3+)											
immonium			X				X	X			X
a	X		X				X	X			
a*	X		X				X				
a0			X				X				
b	X	X	X	X	X	X	X	X			X
b*	X	X	X	X	X	X	X	X			X
b0		X	X	X	X	X	X	X			X
c									X	X	
x											
y	X	X	X	X	X	X	X	X	X	X	X
y*	X	X		X	X	X	X				X
y0		X		X	X	X	X				X
z								X			
yb							X	X			X
ya							X	X			X
y must be significant											
y must be highest score											
z+1									X	X	
d							X				
v							X				
w							X				
z+2									X	X	
Minimum mass											
Max mass	700.000	700.000	700.000	700.000	700.000	700.000	700.000	700.000	700.000	700.000	700.000
		Delete Edit	Delete Edit	Delete Edit	Delete Edit	Delete Edit	Delete Edit	Delete Edit	Delete Edit	Delete Edit	Delete Edit

The INSTRUMENT search parameter is used to select the set of ion series used for scoring MS/MS matches.

File format (fragmentation_rules)

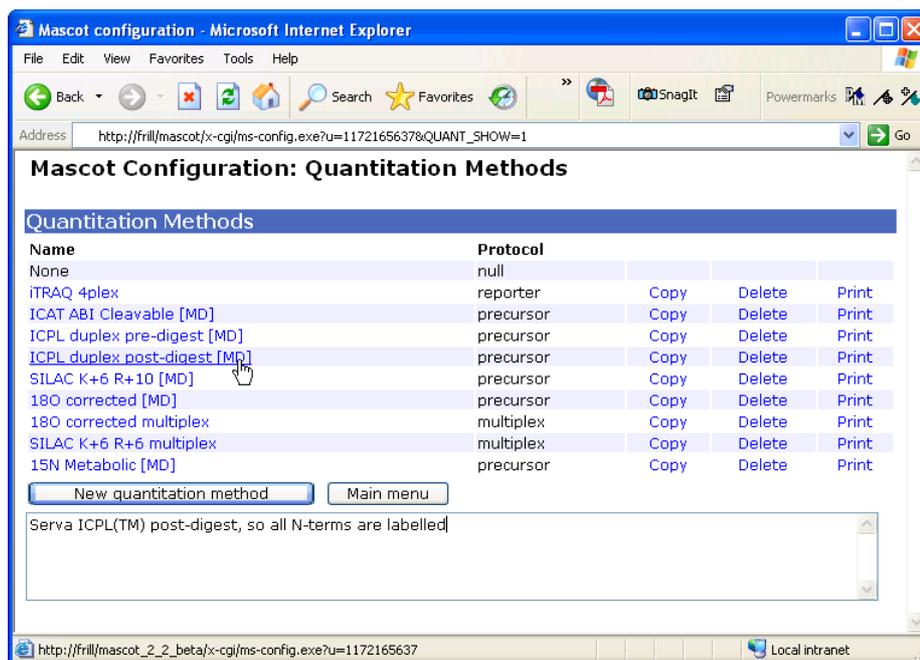
Each instrument is defined by a block of lines. Blocks are delimited from one another by a line containing an asterisk.

The first line of each block must start with the Title: keyword, followed by a text string that is used to identify the instrument in forms and reports. The definition should be short and self-explanatory. It should only include alphanumeric characters and hyphens. The following lines start with an integer, each of which represents an ion series or a rule to be included in the definition. Refer to the file header for a list of available integers. Anything following a hash (#) symbol is treated as a comment.

A block can also specify mass range limits for internal ions. The default range is 0 to 700 Da, and could be changed as in this example:

```
title:MALDI-QIT-TOF
1 # singly charged
4 # immonium
5 # a series
6 # a - NH3 if a significant and fragment includes RKNQ
7 # a - H2O if a significant and fragment includes STED
8 # b series
9 # b - NH3 if b significant and fragment includes RKNQ
10 # b - H2O if b significant and fragment includes STED
13 # y series
14 # y - NH3 if y significant and fragment includes RKNQ
15 # y - H2O if y significant and fragment includes STED
17 # internal yb < 700 Da
18 # internal ya < 700 Da
minInternalMass 200
maxInternalMass 1000
*
```

Quantitation



A detailed description of quantitation methods, the relevant Configuration Editor pages, and the underlying file, (*quantitation.xml*), is contained in the HTML help pages. Choose Help from the Mascot main menu bar and then choose Quantitation.

Database Maintenance

Most sections in the main configuration file, *mascot.dat*, relate to sequence database setup. Database Maintenance, which provides an interface to these sections, is fully described in Chapter 5 and in the HTML help pages.

A few sections of *mascot.dat* have no interface in the Configuration Editor, and the only way to make changes is to edit *mascot.dat*. These sections are Processors, Taxonomy_n, Cluster, UniGene, and Cron

mascot.dat

Note to Windows users

Windows users should note that the path delimiters used in *mascot.dat* must always be forward slashes, never the backward slashes used at the command prompt. If sequence database files are not on a local disk drive, the remote drive must be mapped to a local drive letter. UNC path specifications cannot be used. Finally, spaces are not allowed in file or directory names. Hence:

```
C:/InetPub/mascot/config/mascot.dat      correct ✓
C:\InetPub\mascot\config\mascot.dat      wrong ✗
\\matrix_nt_01\InetPub\mascot\config\mascot.dat  wrong ✗
//matrix_nt_01/InetPub/mascot/config/mascot.dat  wrong ✗
```

General

mascot.dat is divided into sections. Each section starts with a unique keyword and ends with the keyword 'end'.

Comments and blank lines can be used freely. A line which starts with the # character (pound in the US, hash in Europe) is a comment line.

Databases

The Databases section describes the sequence databases:

```
Databases
MSDB   C:/InetPub/Mascot/sequence/MSDB/current/MSDB_*.fasta  ↷
      ↵   AA 312942   0 1 1 1 1 1 17 18 19 7
NCBIInr C:/InetPub/Mascot/sequence/NCBIInr/current/NCBIInr_*.fasta  ↷
      ↵   AA 410000   1 1 1 1 0 0 6 7 0 1
.
.
.
end
```

Each line defines a database using the following 14 parameters:

1. Name: Each database must have a unique name. Ideally, the name should be short and descriptive. Note that these names are case sensitive, and much confusion can be caused by creating (say) *Sprot* and *SPROT*. The name does not need to be the same as or even similar to the filename of the actual FASTA file. Allowed characters are alphanumeric and `_-.$%&() []`

2. Path: FASTA database files must be available locally. Mascot creates its compressed files in the same directory as the original FASTA file. The location of the FASTA file is defined in the Path field. This must be the fully qualified path to the FASTA file, with a wild card in the filename for the purposes described earlier. The delimiters between directories must always be forward slashes, even if Mascot is running on a Windows system.

3. AA / NA: AA for an amino acid (protein) database and NA for a nucleic acid (DNA) database.

4. Obsolete: This parameter used to contain the approximate number of entries (sequences) in the database, used for progress reports during a search. The value is now just a place holder.

5. Obsolete: This parameter used to contain a unique identification number. The value is now just a place holder.

6. Mem map: Flag to indicate whether the database file should be memory mapped (1) or not (0). Database files should always be memory mapped. Unlike memory locking, this does not consume physical RAM.

7. Obsolete: This parameter (Blocks) must always be set to 1.

8. Threads: A Mascot search can use multiple threads. If you are running in cluster mode, 'Threads' must be set to 1. Otherwise, specify the number of threads to be used during searches.

9. Mem lock: Flag to indicate whether a memory mapped database file should be locked in memory (1) or not (0). This setting is only relevant if column 6 contains a 1.

Memory mapped files can be locked in memory, but only if the computer has sufficient RAM. Having a database locked in memory means that it can never be swapped out to disk, ensuring maximum possible search speed. Of course, there also needs to be sufficient RAM for the operating system, (Windows consumes approximately 60 Mb), anything from tens to hundreds of Mb for each Mascot search, and space for any other applications which might be running.

If you try to lock databases into RAM when there isn't room, this will not be a major problem. The locking will fail, generate an error message, and Mascot will carry on regardless. A more serious problem is when there is just sufficient RAM to lock the databases, but none left over for searches or other applications. In this case, the whole system will slow down and the hard disk will be observed to be "thrashing". Eventually, the system is likely to hang or crash.

10. Local ref file: Flag to indicate whether a local reference file is available (1) or not (0). For certain databases (MSDB, OWL, and

SwissProt), it is possible to have a local reference file, from which full text information can be taken for a 'Protein View' report.

11. AccessionParseRule: Index of the regular expression in the PARSE section that can be used to parse an accession string from a FASTA file title line.

12. DescriptionParseRule: Index of the regular expression in the PARSE section that can be used to parse a description string from a FASTA file title line.

13. AccessionRefParseRule: Index of the regular expression in the PARSE section that can be used to parse an accession string from a local full text reference file. If there is no local reference file, this value is ignored and can be set to 0.

14. Taxonomy: Index of the taxonomy rule block to be used to parse taxonomy information. If taxonomy information is not available, or is not to be used, this value should be set to 0.

PARSE

The PARSE section contains Basic Regular Expressions used to extract strings from various files.

```

PARSE
#For OWL 3.1r2 accession e.g. >owl||100K_RAT This is a protein
RULE_1 ">owl[^\ ]*\|([^\ ]*\)"
#
#For OWL 3.1r2 description e.g. >owl||100K_RAT This is a protein
RULE_2 ">owl[^\ ]*\|[^\ ]*\|([^\ ]*\)"
#
#For OWL 3.1 and later .ref file e.g. >P1;100K_RAT
RULE_3 ">[A-Z][1-9];\|([^\ ]*\)"
.
.
.
end

```

The syntax of a standard Basic Regular Expression (BRE) is described in Appendix A. Rules defined in this section are referred to by means of their index number in two sections: Databases and WWW.

The string to be extracted is defined within backslashed parentheses. RULE_1, for example, looks for the text >owl followed by any number of characters which are not a space up to the final pipe symbol. The string to be extracted is any character following the pipe symbol up to the next space.

This rule would successfully extract the accession string 100K_RAT from the following FASTA title lines:

```
>owl||100K_RAT This is a protein ...
>owl|S03653|100K_RAT This is a protein ...
>owl|100K_RAT      This is a protein ...
>owl|gi986574|S036594|100K_RAT This is a protein ...
>owl||100K_RAT
>owl&pussycat||100K_RAT This is a protein ...
```

It would fail in the following cases, which would indicate that a different rule was required:

```
>owl 100K_RAT This is a protein ...
>Owl|S03653|100K_RAT This is a protein ...
>owl|100K_RAT|This is a protein ...
>owl ||100K_RAT This is a protein ...
>owl|| 100K_RAT
>pussycat&owl||100K_RAT This is a protein ...
```

If you are not familiar with regular expressions, use the information in Appendix A to understand how the pre-defined rules in *mascot.dat* work.

A mistake in a rule called from the databases section will prevent Mascot from using the database concerned. Always use the database maintenance utility to test a new database definition before bringing it on-line.

WWW

The WWW section defines where CGI scripts look for the information needed to compile a results report.

At least one line is required for each database, to define the source from which the sequence string of a database entry can be obtained. A second line can optionally define the source from which the full text report of an entry can be obtained. The syntax is very similar in both cases, independent of whether the information originates locally or on a remote system.

Sequence strings can always be retrieved locally, because the FASTA file must be present on a local disk. The Mascot utility *ms-getseq.exe* is normally used to retrieve a sequence string.

Full text for an entry may or may not be available locally. If it is available locally and the database has been defined as including a ref file, (Column 10 in the Database section of *mascot.dat*), *ms-getseq.exe* can be used to retrieve the full text. Otherwise, an application must be identified which can accept an accession string and return the report text in a parseable format. An example of a suitable application would be the SRS system from EBI:

<http://srs.ebi.ac.uk/about.html>

Each line in the WWW section contains 5 columns:

```
WWW
OWL_SEQ      "8"  "localhost"          "80"  "/usr/local/mascot/  ↗
↳ x-cgi/ms-getseq.exe OWL #ACCESSION# seq"
OWL_REP      "9"  "localhost"          "80"  "/usr/local/mascot/  ↗
↳ x-cgi/ms-getseq.exe OWL #ACCESSION# all"
NCBIInr_SEQ  "8"  "localhost"          "80"  "/usr/local/mascot/  ↗
↳ x-cgi/ms-getseq.exe NCBIInr #ACCESSION# seq"
NCBIInr_REP  "10" "www.ncbi.nlm.nih.gov" "80"  "/htbin-post/        ↗
↳ Entrez/query?db=p&form=6&uid=#ACCESSION#&Dopt=g&html=no"
dbEST_SEQ    "8"  "localhost"          "80"  "/usr/local/mascot/  ↗
↳ x-cgi/ms-getseq.exe dbEST #ACCESSION# seq #FRAME#"
dbEST_REP    "11" "www.ncbi.nlm.nih.gov" "80"  "/htbin-post/        ↗
↳ Entrez/query?db=n&form=6&uid=#ACCESSION#&Dopt=g&html=no"
.
.
.
end
```

1. Identifier: An identifier constructed from the name of the database, an underscore character, and either the keyword SEQ or REP. Thus, OWL_SEQ is the source for the sequence string of an entry in the OWL database. dbEST_REP is the source for the full text of an entry in the dbEST database.

2. Parse rule: The index of a rule in the PARSE section that can be used to extract the information required. Note that the rule for parsing a sequence string from *ms-getseq.exe* the same for all databases.

3. Host: The information source. For *ms-getseq.exe* or a similar local executable, this column should contain *localhost*. For a remote source, or a local source that will be queried as a CGI application, enter the hostname. (NB the word *localhost* is used to determine whether the application is a command line executable or a CGI application. If you want to specify a CGI application on the local server, just specify the hostname in some other way, for example *127.0.0.1*).

4. Port: The port number. This should be left at 80 unless another value is required to access a web server operating on a non-default port.

5. Path: A string containing the path to the executable and parameters, some of which are variables.

In the case of a command line executable, the parameters will generally be delimited by spaces. In the case of a CGI application, the parameters must be delimited from the executable by a question mark, and there must be no spaces within the parameter string. In general, spaces in URL's must be replaced by plus symbols, and non-alphanumeric characters should be URL encoded using the %nn notation.

A reminder to Windows users: Do not use backslashes as path delimiters, because these will be interpreted as escape characters.

Most parameters are entered as literal strings, with two exceptions: #ACCESSION# is a place holder that will be replaced by an actual accession string, #FRAME# is a place holder that will be replaced by the number of the reading frame used to translate a nucleic acid sequence. Obviously, this last parameter is only used with NA databases.

The syntax for calling *ms-getseq.exe* is described in Chapter 7. In the examples shown above, full text reports for OWL are taken from a local reference file while full text reports for NCBIInr and dbEST are sourced from the public NCBI web site.

The syntax for linking to NCBI Entrez is described here:

http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

The syntax for calling an external SRS server can be found here:

<http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+docoPage+-newId>

If an SRS server is run locally, a command line interface is also available.

Processors

If the number of processors available is the same as the number licensed, then nothing is required in this section. You can still specify particular processors for particular jobs, but this may have a negative impact on system performance.

If the number of processors available is greater than the number licensed, then nothing is required in this section, but you may wish to specify values.

Processor numbers generally start at 0, but see your computer documentation. Separate processor values with a comma

```

PROCESSORS
ProcessorSet=1,2
OWL=1,2,1,2
end

```

The `ProcessorSet=` line specifies the complete set of processors to be used. The number in this list must be less than or equal to the number licensed, or the system will not run.

Following this, the processors to be used for each database are specified. These numbers must be a subset of the numbers in the `ProcessorSet`, and there must be the same number of values as the number of threads specified earlier in the database section.

This section must be after the Databases section in *mascot.dat*, and `ProcessorSet=` must come before the other entries in this section.

For Intel Pentium 4 Xeon processors, which support hyperthreading, there are restrictions on specifying processors. For further details, refer to the installation notes in Chapter 2 (Unix) or Chapter 3 (Windows).

Taxonomy

The syntax of the taxonomy blocks is fully described in Chapter 9.

Cluster

The syntax of the cluster block is fully described in Chapter 11.

UniGene

UniGene is an index created by automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters, (<http://www.ncbi.nlm.nih.gov/UniGene/>). Each UniGene cluster is a list of the GenBank sequences, including EST's, which represent a unique gene. It is not an attempt to produce a consensus sequence. UniGene can be used to simplify the results of a Mascot search of dbEST.

An index file must be downloaded for each species of interest. For each species, the fully qualified path to the index file is associated with the species name:

```

UniGene
human C:/Inetpub/MASCOT/unigene/human/current/Hs.data
mouse C:/Inetpub/MASCOT/unigene/mouse/current/Mm.data
mosquito C:/Inetpub/MASCOT/unigene/mosquito/current/Aga.data

```

To add a UniGene report option to Mascot for a particular sequence database, add a line containing the name of the database followed by a list of the available species names:

```
EST_human human
EST_mouse mouse
EST_others mosquito
end
```

Options

The Options section is used for miscellaneous parameters, which are listed here in alphabetical order. If a parameter is shown with argument(s), these are the default(s) that apply if the parameter is missing.

```
AutoSelectCharge 0
```

Controls how MS/MS queries are treated when the CHARGE parameter specifies more than one charge state (e.g. 1+, 2+, and 3+). This is usually because no charge information was available for a query, so the search form defaults applied.

If set to 0, a query is generated for each charge state and these queries are searched and reported independently. This is the default setting because this was the behaviour in earlier versions of Mascot.

If set to 1, each charge state will be searched, but only the charge state that gets the highest scoring match is saved to the result file and reported. This is the recommended setting

Note that this switch only applies to MS/MS queries, (including tags). Independent queries are always generated if multiple charge states are specified for molecular mass queries.

```
CentroidWidth 0.25
```

The width in Daltons of the sliding window used for re-centroiding profile data.

```
EmailErrorsEnabled 0
EmailFromTextName
EmailFromUser
EmailPassword
EmailProfile
EmailService
EmailTimeOutPeriod 120
```

```
EmailUsersEnabled 0
ErrMessageEmailTo
MailTempFile C:/TEMP/MXXXXXX
MailTransport 0
MonitorEmailCheckFreq 300
sendmailPath
```

Mascot can be configured to use email for two purposes:

1. When the search engine executes as a CGI application, email can be used to send the results of a search to a user who accidentally or deliberately disconnected before the search was complete. This facility can be enabled by setting `EmailUsersEnabled` to 1 or disabled by setting it to 0.
2. Serious error messages can be emailed to an administrator. This facility can be enabled by setting `EmailErrorsEnabled` to 1 or disabled by setting it to 0. Error messages that are considered serious are identified in the file `errors.html`. This file can be found in the root directory of the installation CD-ROM, and is displayed by clicking on the link 'Error message descriptions' at the top of the database status page.

A number of parameters are used to define how email should be sent:

`MailTransport` should be set to one of the following values:

- 0 for CMC
- 1 for MAPI
- 2 for sendmail
- 3 for Blat

`EmailService` is the service name (CMC only)

`EmailPassword` is the password (if any) required to log onto MAPI or CMC

`EmailProfile` is the MAPI profile name

`sendmailPath` is the path to sendmail, or an equivalent.

`EmailFromUser` is the name which will appear in the 'From' field of the email message.

`EmailFromTextName` will appear in the 'Title' field of the message.

If `EmailUsersEnabled` is set to 1, search results will be emailed to a user if their web browser does not respond within the number of seconds specified in `EmailTimeOutPeriod` following the completion of a search.

Email messages can be sent in batches at intervals specified by `MonitorEmailCheckFreq` (in seconds). `MailTempFile` is the name of the temporary file used to store email messages until they can be sent.

If `EmailErrorsEnabled` is set to 1, serious error messages will be emailed to `ErrorMessageEmailTo`.

MAPI Configuration (Windows Only)

Set `MailTransport` to 1.

Set the `EmailPassword` to the password (if any) that is required to log onto MAPI.

Set the `EmailProfile` to the profile name used by MAPI. This can be found by opening the Windows Control Panel and clicking on Mail. (Depending on whether you have an 'internet mail only' or a 'corporate or workgroup' installation of MS-Outlook, you will have a list of either account names or profile names to choose from).

Sendmail Configuration (Unix Only)

Set `MailTransport` to 2.

Set the `EmailFromUser` parameter to the name that is required in the 'From' field of the email messages.

Set `EmailFromTextName` as the name of the server that is running mascot. For example setting `EmailFromUser` to `www` and `EmailFromTextName` to `Mascot Server` will result in emails from `www (Mascot Server)`. The From field of the email will be `www@www.your_domain.com`.

Set `sendmailPath` as the path for the sendmail program, e.g. `/usr/lib/sendmail`

Set `MailTempFile` as the name of the file used to store email messages until they can be sent (must be the path followed by a filename in the form: `MXXXXXX`). This will create temporary files that begin with `M` followed by a unique number. Typically this parameter will be `/var/tmp/MXXXXXX`.

Blat Configuration (Windows only)

Blat is a free, easily installed mail program for Windows. For more information, visit:

<http://www.blat.net/>

Set `MailTransport` to 3.

Set the `EmailUserFrom` parameter to the name that is required in the 'From' field of the email messages.

Set `EmailFromTextName` as the name of the server that is running mascot. For example setting `EmailUserFrom` to `www` and `EmailFromTextName` to `Mascot Server` will result in emails from `www`

(Mascot Server). The From field of the email will be `www@www.your_domain.com`.

Set `sendmailPath` as the fully qualified path (including drive letter) for the Blat program.

Set `MailTempFile` as the name of the file used to store email messages until they can be sent (must be in the form `path/MXXXXXX`). This will create a new temp file where the first letter will be an M and the next 6 characters will make up a unique number. Typically this parameter will be `c:/temp/MXXXXXX`

```
ErrorLogFile ../logs/errorlog.txt
GetSeqJobIdFile ../data/getseq.job
InterFileBasePath c:/inetpub/mascot/data
InterFileRelPath ../data
MascotCmdLine ../cgi/nph-mascot.exe
MascotControlFile ../data/mascot.control
MascotJobIdFile ../data/mascot.job
MonitorLogFile ../logs/monitor.log
SearchLogFile ../logs/searches.log
TestDirectory ../data/test
UniqueJobStartNumber 001234
```

These entries determine local paths (not URL's). `ErrorLogFile`, `MascotCmdLine`, `MonitorLogFile`, `SearchLogFile`, and `TestDirectory` are self-explanatory.

`GetSeqJobIdFile` contains the next available job number for the `ms-getseq.exe` utility. These numbers wrap around at 999 and do not appear in the search logs. If this file is deleted, the next job number will be reset to 1 and a new `jobId` file created automatically

Mascot output files are written to a path given by:

```
InterFileBasePath/InterFileRelPath/yyyymdd/Fnnnnnn.dat
```

Where `yyyymdd` is the current ISO date, and `nnnnnn` is a sequential job number with a minimum of 6 digits. The path is split into a base path and a relative path as seen by the CGI scripts so that the search engine can pass a file path to (say) `master_results.pl` as:

```
InterFileRelPath/yyyymdd/Fnnnnnn.dat
```

`TestDirectory` contains the input files used by Monitor to test new sequence databases.

`MascotControlFile` contains critical internal parameters. This file must be memory mapped and locked to provide interprocess communication between different Mascot components.

MascotJobIdFile contains the next available job number. If this file is deleted, the next job number will be initialised to the value given by UniqueJobStartNumber, and a new jobId file created automatically. NB UniqueJobStartNumber must never be set lower than 1000.

ErrTolMaxAccessions 0

The maximum number of database entries allowed for a manual error tolerant search. Default is 0, meaning no limit.

FeatureTableLength 30000

If a nucleic acid sequence is longer than 30000 bases, the protein view report will automatically switch to feature table mode and output the matches as a GenBank feature table. The threshold for switching to feature table mode can be altered using the parameter FeatureTableLength in the Options section of mascot.dat or by appending `_featuretablelength=X` to the protein view URL, where X is the length in bases.

FeatureTableMinScore

By default, only matches with significant scores ($p < 0.05$) are output. A different score threshold can be specified using the parameter FeatureTableMinScore in the Options section of mascot.dat or by appending `_featuretableminscore=X` to the protein view URL, where X is the score threshold.

ForkForUnixApache 0

If a user presses 'Stop' or goes to another page in their browser when a search is running, then the intended behaviour is that the search should continue, and the user be emailed with their results. However, when running some versions of Apache, the search is terminated by Apache when the connection to the browser is lost. To stop this from happening, set this value to 1. Setting this parameter to 1 with other servers can cause problems, so only use this setting if necessary. (When set to 1, the result is that `nph-mascot.exe` ignores PIPE signals, does a fork, the parent exits and the child then ignores HUP signals).

FormVersion 1.01

Mascot users may save search forms off-line, or submit searches using scripts or private forms. When the search engine is upgraded, there is the possibility that old scripts or forms may contain invalid or obsolete

parameters. If a search is submitted to Mascot without a version number, or if the version number is lower than that specified by `FormVersion`, a warning will be included in the results file and in the master results report.

`GetSeqJobIdFile` see `ErrorLogFile`

`ICATQuantitationMethod`

For backward compatibility, if a search is submitted from an old client with `ICAT=ON`, then the specified quantitation method will be used.

`IgnoreDupeAccessions` `EST_others`

A comma separated list of database names. For any database in this list, don't check for duplicate accession numbers when creating the compressed files. A database should only be added to this list if it has a very large number of sequence which may causes the system to run out of memory when creating the compressed files.

`IgnoreIonsScoreBelow` `0.0`

When a report is generated, any ions score lower than this value will be set to zero and ignored. The parameter is a floating point number, default 0.0. Values greater than 0 and less than 1 act as an expect value threshold, and the scores for any peptide matches with higher expect values are set to 0. This global default can be over-ridden on an individual report URL by appending `&_ignoreionsscorebelow=X`, where X is the cut-off value.

`IntensitySigFigs` `2`

The precision of intensity values written to the result file.

`InterFileBasePath` see `ErrorLogFile`

`InterFileRelPath` see `ErrorLogFile`

`IonsDecimalPlaces` `2`

Mascot calculates all masses to an accuracy of 1/65535 Daltons. The number of decimal places used to display fragment ion masses in reports can be altered by changing this value.

IteratePMFIntensities 1

Set this option to 0 to prevent selection of PMF values on the basis of their intensity.

LastQueryAscFile ../logs/lastquery.asc

SaveEveryLastQueryAsc 1

SaveLastQueryAsc 0

SaveLastQueryAsc is a flag which controls whether the most recent input file to Mascot (i.e. the MIME format file containing MS data and search parameters) should be saved to disk (1) or not (0). This can be a useful debugging tool when writing scripts or forms to submit searches to Mascot. If SaveLastQueryAsc is set to 1, the name of the file is determined by LastQueryAscFile. Each new search over-writes this file. NB LastQueryAscFile is a disk path, not a URL.

An additional debugging tool is provided by SaveEveryLastQueryAsc. If set to 1, the Mascot input file will be saved for any search that fails to complete because it generates a fatal error. The name of the output file follows the same naming convention as a normal Mascot result file, except for the additional suffix `.inp`. If a search goes to completion, this file is deleted as soon as the normal output file has been written to disk.

LogoImageFile ../images/88x31_logo_white.gif

This is the URL of the Matrix Science logo, used at the top of a search progress report. You can customise this by substituting the URL of your own logo. For optimum appearance, the image should be 88 pixels wide and 31 pixels high.

MailTempFile see EmailErrorsEnabled

MailTransport see EmailErrorsEnabled

MascotCmdLine see ErrorLogFile

MascotControlFile see ErrorLogFile

MascotJobIdFile see ErrorLogFile

MassDecimalPlaces 2

Mascot calculates all masses to an accuracy of 1/65535 Daltons. The number of decimal places used to display peptide mass values in reports can be altered by changing this value.

MaxAccessionLen

Obsolete

MaxConcurrentSearches 10

This parameter limits the maximum number of concurrent searches so as to avoid overloading the Mascot server. Default is 10

MaxDescriptionLen 100

Description text parsed from the FASTA title line will be truncated at this number of characters. (Note: There is no need to recompress a database if this parameter is changed).

MaxEtVarMods 2

The maximum number of variable mods allowed in the first pass of an automated error tolerant search (global default, can be over-ridden for a group in security)

MaxNumPeptides

The maximum number of peptides that can be expected from the enzymatic digest of a single entry. The default is $\text{MaxSequenceLen} / 4$

MaxPepNumVarMods 5

The maximum number of variable mods allowed for a PMF (global default, can be over-ridden for a group in security)

MaxQueries 10000

The maximum number of MS/MS spectra allowed in a single search. Note that the maximum number of mass values in a PMF is hard-coded to 1000

MaxSearchesPerUser 0

Sets the maximum number of concurrent searches from a single IP address. A value of 0 means no limit. (global default, can be over-ridden for a group in security)

MaxSequenceLen 50000

The maximum length of a database entry in characters, (bases for NA or residues for AA). The default is 50,000. The length of the longest sequence in a database can be found in the *.stats file, created by Mascot Monitor when the database is compressed. The larger the value of MaxSequenceLen, the more memory mascot uses. So, if you need to increase it, make it just a little greater than the length of the longest sequence. On a 32 bit system, try not to exceed 3 million, because searches may run slower than normal. If you are trying to search an assembled genome, you might want to consider searching shorter sequences instead, such as a database of the contigs.

MaxVarMods 9

The maximum number of variable mods allowed for an MIS search (global default, can be over-ridden for a group in security)

MinPepLenInPepSummary 6

In a Peptide Summary report, two proteins are reported as distinct matches if the peptide matches to one protein are not identical to or a sub-set of the peptide matches to the other protein. Since matches to very short peptides are usually random, peptides shorter than MinPepLenInPepSummary are not considered in this comparison. Default value is 6.

MinPepLenInSearch 5

Peptides shorter than MinPepLenInSearch are rejected during the search. Matches to very short peptides are meaningless because a 2-mer or 3-mer can occur in almost every entry in a database. If such matches are allowed in the peptides section, it can cause serious bloating of the result file. Default value is 5.

MonitorEmailCheckFreq see EmailErrorsEnabled

MonitorLogFile see ErrorLogFile

MonitorPidFile monitor.pid

The name for the file that holds the process ID number for *ms-monitor.exe*. Default is *monitor.pid*.

MonitorTestTimeout 1200

A time-out can be applied to the test searches used to validate a new database. If the test search on a new database does not produce a valid result within the number of seconds specified by `MonitorTestTimeout`, the problem is assumed to be with the new database, and the exchange process is halted.

MoveOldDbToOldDir 1

After a successful database swap, the old Fasta file and old reference file (if any) are moved to the `../old` directory unless this parameter is present and set to 0

Mudpit 1000

Obsolete, see `MudpitSwitch`

MudpitSwitch 0.001

Mascot has two ways to calculate protein scores in a Peptide or Select summary report. Standard scoring is used when the ratio between the number of queries and the number of database entries, (after any taxonomy filter), is small. The standard score is the sum of the ion scores after excluding duplicate matches and applying a small correction. Protein score calculation switches to large search mode when the ratio between the number of queries and the number of database entries, (after any taxonomy filter), exceeds the value specified by `MudpitSwitch`. Only those ions scores that exceed one or both significance thresholds contribute to the score, so that low scoring, random matches have no effect. The global default can also be over-ridden on an individual report URL by appending `&_server_mudpit_switch=X`, where X is the ratio between the number of queries and the number of database entries, (after any taxonomy filter).

```
NoResultsScript ../cgi/master_results.pl
ResultsFullURL
ResultsPerlScript ../cgi/master_results.pl
```

These are URL's (not disk paths) for the scripts to be called by the search engine at the completion of a search. A successful search calls `ResultsPerlScript`, while a search that didn't find any hits calls `NoResultsScript`. In the scripts supplied with Mascot, both eventualities are handled by a single script: `master_results.pl`.

The `ResultsFullURL` entry is used when results are emailed to a user. Since the email will probably be received on another system, the URL used needs to have the full URL including the Web server hostname.

`NTIUserGroup` `Guests`
`NTMonitorGroup` `Administrators`

Under Windows, the Mascot service is generally run using the 'Local System' account. It has to create, write and read the memory mapped files. The CGI scripts (such as `nph-mascot.exe`) are run by the Web server, and will be run using a different user name with different permissions from the service. These programs also need to be able to read and write to these files. For example, with the Microsoft Web server (IIS), a new user with the name `IUSR_<name_of_pc>` is created when the server is installed, and the scripts are run using this user name. The installation program sets these values appropriately. Other Web servers may use different user names, with different permissions.

`NTMonitorGroup` is used to define one of the groups that the administrator belongs to. In most cases, this can be left at the default of *Administrators*

`NTIUserGroup` is the name of one of the groups that the user name used to run the CGI scripts belongs to. With the Microsoft Personal Web server (running on Windows NT Workstation), this will be *Guests*. With the Microsoft Web server (IIS) running under Windows NT / 2000 Server configured as a domain controller, this will need to be changed to *Domain Users*. (The installation program sets these values appropriately, so it should not need to be done manually).

For a different Web server, check the documentation that comes with the server to find out which user name is used for running scripts, then from the start menu, choose, Programs, administrative tools (common), and User Manager. Double click on the user name, and press the groups button to find out which groups this user name belongs to. This is the name to put in `mascot.dat` for `NTIUserGroup`.

Failure to put the correct group name will generally result in one of two error messages:

```
Failed to open memory mapped file <filename>. ␣  
␣ Error: access denied
```

or

```
Failed to create memory map for <filename>. ␣  
␣ Error Access denied
```

After changing either of these entries, the Mascot service will need to be stopped, (from the start menu, choose *Programs; Mascot; config;*

Stop Mascot service). All compressed database files must be deleted. Then the Mascot service can be re-started (*Programs; Mascot; config; Start Mascot service*).

PrecursorCutOut -1,-1

The precursor peak can often have very high intensity relative to the fragment peaks, which may give rise to spurious fragment ion matches. It is usually best if the precursor is removed before the search.

With the default arguments of -1,-1, a smart filter is created. This removes peaks within the fragment ion tolerance window about each of the precursor isotope peaks. The number of isotopes is assumed to be as follows:

Mr	Number
< 1000	3
1000 - 1999	4
2000 - 2999	5
3000 - 3999	6
4000 - 4999	7
5000 - 5999	8
6000 - 6999	9
> 7000	10

So, if the precursor m/z was 800, the charge was 2, and fragment ion tolerance was +/- 0.1 Da, the filter would remove 4 notches of width

```
m/z 800.0 +/- 0.1
m/z 800.5 +/- 0.1
m/z 801.0 +/- 0.1
m/z 801.5 +/- 0.1
```

At first sight, this may seem a strange mix of m/z and Da. The reason is that we need to avoid matches from 1+ fragment ions, whatever the charge on the precursor.

If the arguments are anything other than -1,-1, a single notch is used where the first argument is the mass offset of the beginning of the notch and the second value is the mass offset of the end of the notch. For the precursor in the last example, if the arguments were -1,4 then the notch would run from m/z 799.5 to m/z 802.0. However, if the precursor charge was 1, then the notch would be from m/z 799 to m/z 804.

The mascot.dat setting can be over-ridden in a search by using the search parameter CUTOUT. Note that the peaks removed by this filter

are not recorded in the result file, so cannot be recovered by changing this parameter in a repeat search.

ProteinsInResultsFile 2

Determines the number of protein title lines saved to each results file.

- 1 As in Mascot 1.7 and earlier, only proteins that appear in the Summary section will appear in the Proteins section
- 2 Include proteins with at least one top ranking peptide match to a peptide of length greater than `MinPepLengthInPepSummary`
- 3 Include all proteins

proxy_password
proxy_server
proxy_username

These entries support a proxy server between the Mascot server and the outside world. A typical entry might be

```
proxy_server http://our-cache:3128
```

If there is no `proxy_server` entry, scripts will look for proxy information in the server environment. The `proxy_username` and `proxy_password` parameters are only required if the proxy server requires authentication. Remote host authentication should be included directly in the URLs specified in `mascot.dat`. e.g. `http://username:password@hostname/`

RemoveOldIndexFiles 1

After a successful database swap, the compressed files in the current directory are deleted unless this parameter is present and set to 0

ReportNumberChoices 5,10,20,30,40,50

If present, this list will define the choices provided in the search form 'Report top' drop down list.

RequireBoldRed 0

If this flag is set to 1, only protein matches which have one or more 'bold red' peptide matches will be listed in a peptide summary report. That is, proteins that include at least one top ranking peptide match that has not

already appeared in the report. This global default can be overridden on an individual report URL by appending `&_requireboldred=X`, where X is 0 or 1.

ResultsFileFormatVersion

If present, and the argument is 2.1, the result file format will be “2.1 compatible”. That is, no xml sections. No other arguments are supported at this time.

ResultsFullURL see NoResultsScript

ResultsPerlScript see NoResultsScript

ReviewColWidths 7,8,8,27,30,100,32,25,6,13,2,4,6,16,7

This sets the widths of the columns in ms-review.exe.

SaveEveryLastQueryAsc see LastQueryAscFile

SaveLastQueryAsc see LastQueryAscFile

ScoreThresholdForAuto

Deprecated, use SigThreshold.

SearchControlLifetime 7200

SearchControlSaveE 0

Obsolete.

SearchLogFile see ErrorLogFile

SendmailPath see EmailErrorsEnabled

SelectSwitch 1000

If the number of queries in an MS/MS search is less than or equal to this number, the default report is the Peptide Summary. If it is greater than this number, the default report is the Select Summary.

SeparateLockMem 0

Only required if the total amount of memory to be locked is greater than 2Gb (or lower if some system limit is set). Setting this value to 1 indicates that *ms-monitor.exe* will run a separate program (*ms-lockmem.exe*) that will lock the memory blocks. A value greater than 1 specifies the block size in Mb. For example, if there is a 1.5 Gb *.s00 file, and this parameter is set to 750, then two instances of *ms-lockmem.exe* will be run.

ShowAllFromErrorTolerant 0

Standard behaviour for the result report of a manual error tolerant search is to show only those matches that satisfy two criteria: (i) the score must be at least as high as the match for the same query in the original 'parent' search, (ii) the score equals or exceeds the identity threshold for the same query in the original 'parent' search. Setting ShowAllFromErrorTolerant to 1 causes all matches to be displayed. This global default can be overridden on an individual report URL by appending `&_showallfromerrortolerant=X`, where X is 0 or 1.

ShowSubSets 0

If this is set to 1, under each protein match in a peptide summary report, matches to proteins that contain a sub-set of the same peptides will also be listed. This was the default behaviour in version 1.6 and earlier. If this flag is set to 0, which is now the default, the sub-set matches will not be shown. Values between 0 and 1 represent the fraction of the protein score of the primary hit that a subset hit can lose and still be listed. For example, if ShowSubSets is 0.2, and the primary hit has a protein score of 200, sub-set hits with scores of 160 or more will be listed.

If multiple entries contain the full set of peptides, they are all displayed, whatever the setting of this parameter. This global default can be overridden on an individual report URL by appending `&_showsubsets=X`, where X is 0 or 1.

SigThreshold 0.05

Significance threshold used in result reports, default 0.05. Valid range is 1 to 1E-18. This global default can be overridden on an individual report URL by appending `&_sigthreshold=X`, where X is the significance threshold.

SortUnassigned scoredown

In a peptide summary report, peptide matches that are not assigned to protein hits are initially sorted by descending score (scoredown). Alternatives for SortUnassigned are ascending query order (queryup) and descending intensity order (intdown). This global default can be overridden on an individual report URL by appending `&_sortunassigned=X`, where X is scoredown, queryup, or intdown.

SplitDataFileSize 10000000

Large searches are divided into ‘chunks’, and no single chunk can exceed this number of bytes – default 10 Mb. When a search is divided into chunks, protein and peptide match data are no longer written to the summary section of the result file. This means that a Protein summary report cannot be generated.

SplitNumberOfQueries 1000

Large searches are divided into ‘chunks’, and no single chunk can exceed this number of queries – default 1000. When a search is divided into chunks, protein and peptide match data are no longer written to the summary section of the result file. This means that a Protein summary report cannot be generated.

StoreModPermutations 1

If set to 0, only the highest scoring permutation of variable modifications for each unique peptide sequence is retained in the list of the top 10 ions scores. If set to 1, then different permutations of variable modifications are treated as independent matches, creating the possibility that all 10 top ions scores correspond to the same primary sequence. Default is 1.

TaxBrowserURL

(No default). The URL used in reports to retrieve taxonomy information for a Protein View report. By default, this points to the NCBI. If you don’t want to send such queries out to the internet, the URL can be replaced by a call to the `ms-gettaxonomy.exe` utility:

```
TaxBrowserUrl ../x-cgi/ms-  
gettaxonomy.exe?4+#DATABASE#+#ACCESSION#
```

TestDirectory see ErrorLogFile

UniqueJobStartNumber see ErrorLogFile

UnixDirPerm 777

Specify the Unix permissions for the 'daily' result file directories. For example, 775 makes each directory world readable but not writeable. This option provides more fine grained control than UnixWebUserGroup

UnixWebUserGroup

This entry, if present, will restrict access to the files created by *ms-monitor.exe*, and hence improve system security. The UnixWebUserGroup is the number of the group used by the web server to run CGI programs. With Apache, the group name will generally be nobody, and you will need to ascertain the group number from the group file. For other Web servers, check the documentation that comes with the server to find out which user name is used for running CGI programs.

A value of -2 can be used if the same user name is used to run Web server scripts as runs *ms-monitor.exe*. (This is generally only possible under Irix, using capabilities). In this case, The files created by *ms-monitor.exe* will not be world accessible, and 'chown' is not used on the files to change ownership.

Failure to put the correct group name will generally result in one of two error messages:

```
Failed to open memory mapped file <filename>. ㄿ  
  ㄿ Error: access denied
```

or

```
Failed to create memory map for <filename>. ㄿ  
  ㄿ Error Access denied
```

Vmemory -1

Obsolete.

Cron

On Unix systems, cron can be used to automate routine procedures such as the overnight updating of sequence database files. Windows 2000, XP, and 2003 include an equivalent utility called 'Scheduled Tasks'. Windows NT does not include a suitable utility so, as a convenience to Windows NT users, Monitor can be configured to emulate the functionality of cron.

```
Cron
CronEnable 0
```

CronEnable should be set to 1 to enable cron functionality, 0 to disable. The remaining lines in this section simulate a crontab file

```
45 0 1-31 * * perl c:/inetPub/mascot/bin/mirror.pl
30 3 1-23,25-31 * * perl c:/inetPub/mascot/bin/db_update.pl
                        NCBIInr_from_NCBI
30 3 24 * * perl c:/inetPub/mascot/bin/db_update.pl
                        MSDB_from_EBI
end
```

A crontab file consists of lines of six fields each. The fields are separated by spaces or tabs. The first five are integer patterns that specify the following: minute (0-59), hour (0-23), day of the month (1-31), month of the year (1-12), day of the week (0-6 with 0=Sunday). Each of these patterns may be either an asterisk (meaning all legal values) or a list of elements separated by commas.

An element is either a number or two numbers separated by a minus sign (meaning an inclusive range). Note that days may be specified in two different ways (day of the month and day of the week). If both are specified as a list of elements, both are adhered to. For example,

```
0 0 1,15 * 1
```

would run a command on the first and fifteenth of each month, as well as on every Monday. To specify days by only one field, the other field should be set to * (for example, 0 0 * * 1 would run a command only on Mondays).

The sixth field of a line in a crontab file is a string that is executed by the shell (command prompt) at the specified times. The string must be on a single line. The entire string, up to the end of the line, is passed to the command prompt for execution. The part of the string up to the first space must be the fully qualified path to an executable. The remainder of the line will be passed to the command as parameters.

The examples shown here are:

Back-up the previous day's results files at 00:45 every day

Update NCBIInr every day except for the 24th of the month

Update MSDB on the 24th of the month.

Note that the `mirror.pl` script supplied with Mascot is only an example, and will almost certainly require some degree of modification to reflect your local server configuration.

Log files

Mascot maintains several log files, which are described below. When trouble-shooting, it can be useful to inspect the web server log files, also. Errors in Perl scripts, for example, will be appear in the web server error log, not the Mascot error log.

Error Log

All errors are logged to `logs/errorlog.txt`. This is may be the only place to find a fatal error message resulting from a major configuration problem.

Examples of typical error messages are shown below. A comprehensive list of all Mascot error messages can be found in the file `errors.html`, in the root directory of the Mascot CD-ROM.

```
Error [M00088 - Job 2636 - X00123:file-upload] - Thu Mar 11 10:59:30 1999
- Invalid command/mass at line 1 of your query.
Line is where am I?
```

```
Error [M00034 - Job 2638 - X00251:modifications] - Thu Mar 11 10:59:59 1999
- Modification conflict: Both Carbamidomethyl (C) and ↵
  ↵ Carboxymethyl (C) modify the same residue
```

```
Error [M00133 - Job 2639 - X00938:www] - Thu Mar 11 11:00:21 1999
- Peptide mass of -1234 is too small. The minimum mass allowed is 30
```

Searches Log

Every Mascot search is listed in `logs/searches.log`. The Mascot Review utility provides a web browser interface to this file, displaying filtered and sorted listings of searches. Mascot Review is fully described in Chapter 7.

Alternatively, the file can be opened in a spreadsheet program. The file consists of 14 columns, delimited by tabs. Row 1 contains column titles. An example of a single entry is shown below:

```
2633 \t 185 \t NCBIInr \t JSC \t JSC@work \t \t ↵
  ↵ ../data/19990311/F002633.dat \t Thu Mar 11 09:10:36 1999 ↵
  ↵ \t 17 \t User read res \t 1 \t PMF \t Yes \t 192.168.42.4
```

(Tabs indicated by \t for clarity). The individual columns contain the following information:

Column 1: Mascot job number. Job numbers are allocated sequentially, but will appear in the log in the order in which searches are completed. If the submitted search contained an error which prevented the search starting, there will be no entry in `searches.log`, but there should be an entry in `errorlog.txt`.

Column 2: Process ID

Column 3: Sequence Database searched

Column 4: User name. User names are required by the (JavaScript) search forms, but not by the search engine, so this field may be empty. If an entry logs utility program activity, rather than a search, this field contains the name of the utility, e.g. `TESTPARSE` or `GETSEQ`.

Column 5: User email address. User email addresses are required by the (JavaScript) search forms, but not by the search engine, so this field may be empty.

Column 6: Search title. Empty if none supplied.

Column 7: Relative path to Mascot search results file

Column 8: Start time in the format illustrated in the example above.

Column 9: Duration in seconds

Column 10: Completion Status, normally `"User read res"`. If `EmailUsersEnabled` is set to 1, and the user disconnected before the search was complete, this entry would read `"user emailed"`.

Column 11: Job Priority. Not currently implemented

Column 12: Type of search: `PMF`, `SQ`, or `MIS`

Column 13: Enzyme: Either `yes` (if user selected an enzyme) or `no` (if user selected enzyme type `None`).

Column 14: User IP address

Monitor Log

Mascot Monitor activity, such as sequence database exchange, is logged to *logs/monitor.log*. The following extract shows a typical example of the contents:

```

Fri Sep 14 13:50:02 2001 - -----ms-monitor started
Fri Sep 14 13:50:02 2001 - Locked memory for file ../data/mascot.control
Fri Sep 14 13:50:03 2001 - Checking that Mascot Nodes exist      to Loading DB informa-
tion
Fri Sep 14 13:50:04 2001 - Loading DB information                to Started up success-
fully
Fri Sep 14 13:50:05 2001 - MSDB      0 Not in use                to Preparing to run 1st
test
Fri Sep 14 13:50:05 2001 - MSDB      0 Preparing to run 1st test to About to compress
files
Fri Sep 14 13:50:06 2001 - MSDB      0 About to compress files   to Creating compressed
files
Fri Sep 14 13:50:10 2001 - Creating compressed files from C:/INETPUB/MASCOT/sequence/
MSDB/current/MSDB_20010401.fasta
Fri Sep 14 13:50:10 2001 - Creating compress file C:/INETPUB/MASCOT/sequence/MSDB/
current/MSDB_20010401.i00
Fri Sep 14 13:50:10 2001 - Creating compress file C:/INETPUB/MASCOT/sequence/MSDB/
current/MSDB_20010401.s00
Fri Sep 14 13:50:10 2001 - Creating compress file C:/INETPUB/MASCOT/sequence/MSDB/
current/MSDB_20010401.a00
Fri Sep 14 13:50:10 2001 - Creating compress file C:/INETPUB/MASCOT/sequence/MSDB/
current/MSDB_20010401.t00
Fri Sep 14 15:11:42 2001 - MSDB      0 Creating compressed files   to Finished compressing
files
Fri Sep 14 15:11:43 2001 - MSDB      0 Finished compressing files to Running 1st test
Fri Sep 14 15:11:43 2001 - Running system command: ..\cgi\nph-mascot.exe 2
..\data\test\MSDB_20010401.fasta.testedOk 0 < ..\data\test\MSDB.asc
Fri Sep 14 15:11:43 2001 - Running system command: mkdir
C:\INETPUB\MASCOT\data\..\data\20010914
Fri Sep 14 15:13:01 2001 - MSDB      0 Running 1st test           to First test just run
OK
Fri Sep 14 15:13:02 2001 - MSDB      0 First test just run OK     to Waiting for other DB
to end
Fri Sep 14 15:13:03 2001 - MSDB      0 Waiting for other DB to end to Trying to memory map
files
Fri Sep 14 15:13:05 2001 - MSDB      0 Trying to memory map files to Just enabled memory
mapping
Fri Sep 14 15:13:06 2001 - MSDB      0 Just enabled memory mapping to In use
.
.
.

```

IPC Log

In cluster mode (only) an interprocess communication log can be enabled by setting `IPCLogging` (in the cluster section of *mascot.dat*) to 1 or 2. This log can be used to investigate communications errors at the socket level.



Program Reference

Mascot implements a client-server architecture using the HTTP protocol, (web server / web browser). In this mode, the search engine is run by the web server, as a CGI application.

For some applications, such as closed-loop automation, the search engine may be executed as a 'console' or 'command line' application. This Chapter provides the information that is required to write scripts or applications which interface to the Mascot search engine and associated programs.

Mascot Search Engine

The Mascot search engine, *cgi/nph-mascot.exe*, accepts three command line arguments and a MIME format ASCII text file on standard input (STDIN) containing search data and parameters.

```
nph-mascot.exe 1 [-commandline] [-f path]
  [--taskID number] [--sessionID string] < in.asc
```

The first argument is required, and is a digit, between 1 and 4, which determines the mode of operation:

- 1: Normal search; MS/MS data, if any, form part of the MIME format input file
- 2: Monitor test mode 0
- 3: Monitor test mode 1
- 4: Repeat search; the MIME format input file contains a reference to a Mascot results file which may contain MS/MS data

Optional argument `-commandline` is a flag. If present, HTML formatted output is not written to STDOUT.

Optional argument `-f` allows a result file path to be specified. In the absence of this argument, the result file will be written to a daily sub-

directory of *mascot/data* and have the filename *F123456.dat*, where 123456 is an auto-incremented job number.

Optional argument `--taskID` is used to specify a unique numeric identifier. This identifier should be obtained from the SearchControl utility, described later in this chapter. By specifying an identifier, progress reports and search results can be obtained asynchronously from SearchControl.

Optional argument `--sessionID` is used to specify a Mascot security session identifier, (see Chapter 12).

The file piped to STDIN must be a MIME format file containing the search parameters and mass spectrometry data.

Monitor test mode has a different syntax:

```
nph-mascot.exe 2|3 path [number] < in.asc
```

Required argument *path* is the path to a flag file, e.g. *../data/test/SwissProt_51.0.fasta.testedOk* and optional argument *number* is the cluster number. The input file, e.g. *../data/test/SwissProt.asc*, is created automatically from the *OWL.asc* template

The Monitor application must be running before search engine can be invoked. During search execution, warnings, errors, progress reports, etc. are written to standard output (STDOUT). This output is formatted as HTML text for viewing on a web browser. If the search engine is not being executed as a CGI application, the calling application may need to parse the output to remove the HTML tags.

When a search is complete, an HTML string is written to STDOUT, which causes the client browser to invoke the script defined in *mascot.dat* for displaying a results report, (*master_results.pl*). If the search engine is not being executed as a CGI application, the name of the results file can be parsed directly from this string. The output to STDOUT from a successful search will closely resemble the following:

```
(null) 200 OK
Server: (null)
Content-type: text/html
Pragma: no-cache

<HTML>
<HEAD><TITLE>Mascot searching...</TITLE>
<META HTTP-EQUIV="Expires" CONTENT="0">
<META HTTP-EQUIV="Pragma" CONTENT="no-cache">
</HEAD><BODY BGCOLOR=' #FFFFFF' >
<!-- comment here -->
<!-- comment here -->
<!-- comment here -->
<!-- comment here -->
```

```

<!-- comment here -->
<H1><IMG SRC="../images/88x31_logo_white.gif" WIDTH="88" HEIGHT=
  "31" ALIGN="TOP" BORDER="0" NATURALSIZEFLAG="3"> Mascot Search</H1>
Licensed to: Matrix Science In-house NT system.<BR>Not a real form
Finished uploading search details...<BR>
<B>IMPORTANT:</B> If you get disconnected or choose not to wait
  for your search results<BR>DO NOT RESUBMIT THE SEARCH. Your
  results will be sent by email when the search is complete<BR>

Searching...<BR>
.10% complete<BR>
..20% complete<BR>
...30% complete<BR>
....50% complete<BR>
.....60% complete<BR>
.....70% complete<BR>
.....90% complete<BR>
271397 sequences and 86500527 residues checked.<BR>
<SCRIPT LANGUAGE="JavaScript">
<!-- Begin hiding Javascript from old browsers.
if(window.navigator.userAgent.indexOf("MSIE") != -1){
  window.location.replace("../cgi/master_results.pl?file=
  ../data/19990312/F002642.dat");
} else if (window.location.replace == null){
  window.location.assign("../cgi/master_results.pl?file=
  ../data/19990312/F002642.dat");
} else {
  window.location.replace("../cgi/master_results.pl?file=
  ../data/19990312/F002642.dat");
}

// End hiding Javascript from old browsers. -->
</SCRIPT>
<NOSCRIPT>
<A HREF="../cgi/master_results.pl?file=../data/19990312/F002642.dat">
  Click here to see Search Report</A>
</NOSCRIPT>
</BODY></HTML>

```

Monitor

The primary function of Mascot Monitor, *bin/ms-monitor.exe*, is to manage the sequence databases. Monitor must be running in order for

the search engine to execute. Under Unix this runs as a daemon, and under Windows this runs as a service.

Monitor does the following:

1. Creates compressed files from the databases, checking that the FASTA database files are valid; minor errors in the files are reported as warnings, more serious errors stop the databases from being used
2. These files can then be mapped into memory to improve search times
3. Allows swapping and updating of databases without interruption to executing searches. This means that Mascot can be available 24 hours a day, 7 days a week. This allows Mascot to take advantage of daily updates to the major public databases.
4. Deletes old copies of the FASTA databases to stop the disk becoming full; only the most recent copy is kept.
5. Under Windows NT, the monitor service can run scripts that download updates to the databases. Under later versions of Windows, Scheduled Tasks provides a more convenient mechanism, while under Unix, cron would normally be used.
6. Optionally email a system administrator with serious errors requiring immediate attention. Configuration of email settings in the options section of *mascot.dat* is described in Chapter 6.
7. Optionally email users with their results if they didn't wait for them. Configuration of email settings in the options section of *mascot.dat* is described in Chapter 6.

Sequence Of Events When A New Database Is Added

When a new or updated database is added to a directory, the following sequence of events takes place:

1. If the entry in the *mascot.dat* file indicates that there should also be a reference file containing full text entries, Monitor looks for a file with the same name as the new file but with a *.ref* extension instead of *.fasta*. If there is no such file, the swap to the new database stops.
2. Compressed index files are made from the *.fasta* and *.ref* files. For example, the following files would be created for the database *msdb_20000830*:
 - msdb_20000830.a00*
 - msdb_20000830.s00*
 - msdb_20000830.i00*
 - msdb_20000830.t00* (only if there is taxonomy)

These files are a Mascot proprietary format, which is unlikely be useful for other applications.

3. If a serious error occurs while creating these files, then the conversion to the new database stops, an error is put into the error log and (optionally) the error message is emailed to the administrator. Also, if the status screen is shown, the existence of the error is shown on that screen. Searches on the existing database will continue until the problem is resolved.

4. A test search is performed on the new database. The test uses the appropriate file in the `../data/test` directory. If the test is successful, then a file with the name `<database_name>.testedOk` is put into the `../data/test` directory. If the test fails, then an error is put into the error log and (optionally) the error message is emailed to the administrator. Also, if the status screen is shown, the existence of the error is shown on that screen. Searches on the existing database will continue until the problem is resolved.

5. Any new searches submitted by users will now use the new database.

6. When there are no more searches running that use the old database, the files for the old database will be unmapped from memory, and the new files are then mapped into memory.

7. Any files in the `old` directory for the database, which have the same base name as the current files, are deleted.

8. The `.fasta` and `.ref` files for the outgoing database are moved to the `old` directory

9. The compressed index files for the outgoing database are deleted.

Why Memory Map and Lock the FASTA Files?

To speed up the processing of the FASTA files, they should be mapped into memory. This is particularly important:

1. On a PC or similar platform with a low cost disk system, where disk access is synchronous and not very fast
2. On all platforms where simultaneous searches cause access to different parts of the database at the same time

Databases can be configured (in `mascot.dat`) for three operational modes:

1. Without memory mapping. Do not choose this option, it will make searches very slow.

2. Memory mapping the database files, but not locking the memory. This gives the best performance in a system with insufficient memory. When the system gets low on memory, the files are swapped out of memory and no are longer mapped. On most platforms, this will give better performance than simply relying on the system file cache.

3. Memory mapping the files, and locking the memory. This gives the best possible performance, but does require sufficient RAM for the databases, the operating system, Mascot, and any other applications that are to run concurrently with Mascot.

In order to reduce the amount of memory required, and to prevent memory fragmentation, the sequence strings from the FASTA database are saved separately in a number of files that are then memory mapped. The description line(s) are not memory mapped, only an index to the description in the original database. Compared with mapping the original FASTA database, this can reduce memory requirements by more than 30%. Furthermore, the savings for a nucleic acid database are even greater because the files are compressed with a 2:1 ratio.

If Monitor is started from a command or shell prompt, output similar to the following will be obtained. The same output can be viewed in log file, *monitor.log*.

```
-----ms-monitor started
Locked memory for file ../data/mascot.control
Checking that Mascot Nodes exist      to Loading DB information
Loading DB information                to Started up successfully
MSDB  0 Not in use                    to Preparing to run 1st test
MSDB  0 Preparing to run 1st test     to About to compress files
Sprout 0 Not in use                   to Preparing to run 1st test
Sprout 0 Preparing to run 1st test   to About to compress files
MSDB  0 About to compress files       to Compressed files -> cluster
Sprout 0 About to compress files     to Compressed files -> cluster
MSDB  0 Compressed files -> cluster  to Finished compressing files
Sprout 0 Compressed files -> cluster to Finished compressing files
MSDB  0 Finished compressing files   to First test just run OK
Sprout 0 Finished compressing files  to First test just run OK
MSDB  0 First test just run OK       to Waiting for other DB to end
Sprout 0 First test just run OK     to Waiting for other DB to end
MSDB  0 Waiting for other DB to end  to Trying to memory map files
Sprout 0 Waiting for other DB to end to Trying to memory map files
MSDB  0 Trying to memory map files   to Just enabled memory mapping
Sprout 0 Trying to memory map files  to Just enabled memory mapping
MSDB  0 Just enabled memory mapping  to In use
Sprout 0 Just enabled memory mapping to In use
```

N.B. Under Windows, *ms-monitor.exe* must not be started from the command line if it is already running as a service.

Status and error messages from Monitor can also be viewed from a web browser using the Mascot Status application, described below.

GetSeq

GetSeq is a utility for retrieving the sequence, title, or full text of an entry in a database configured for use by Mascot.

The executable, *x-cgi/ms-getseq.exe*, accepts 3 or 4 command line parameters:

1. The name of the database, e.g. OWL. This argument is required.
2. An accession string, e.g. 100K_RAT. This argument is required.
3. One of five keywords: seq, all, len, title or pI. This argument is required, and is explained further below.
4. (Nucleic acid databases only) Frame number between 1 and 6 to retrieve a sequence translated into protein or 0 for the original nucleic acid sequence.

If the keyword `seq` is supplied, the output from GetSeq has the following format:

Content-type: text/plain

```
*MMSARGDFLNYALSLMRSHNDEHSDVLPRLY ... PLYSSKQILKQKLLLAIKTKNFGFV
>100K_RAT 100 KD PROTEIN (EC 6.3.2.-). - RATTUS NORVEGICUS (RAT).
```

The keyword, `all`, is only applicable if a local, full text database is available and configured in *mascot.dat*. In which case, the returned text has a format similar to the following:

Content-type: text/plain

```
*MMSARGDFLNYALSLMRSHNDEHSDVLPRLY ... PLYSSKQILKQKLLLAIKTKNFGFV
>100K_RAT 100 KD PROTEIN (EC 6.3.2.-). - RATTUS NORVEGICUS (RAT).
>P1;100K_RAT
100 KD PROTEIN (EC 6.3.2.-). - RATTUS NORVEGICUS (RAT).
.
.
.
```

```
C;DOMAIN 827 847 PRO-RICH.  
C;BINDING 858 858 UBIQUITIN (BY SIMILARITY).  
C;Keywords: UBIQUITIN CONJUGATION; LIGASE.
```

In all cases, the first line is a content-type specifier, followed by a blank line.

For `seq` and `all` there is then an asterisk followed by the unformatted sequence in one letter code. The next line is identical to the FASTA title line, beginning with a right angle bracket.

In the case of a full text report, this is followed by the raw text entry, as retrieved from the sequence database full text file.

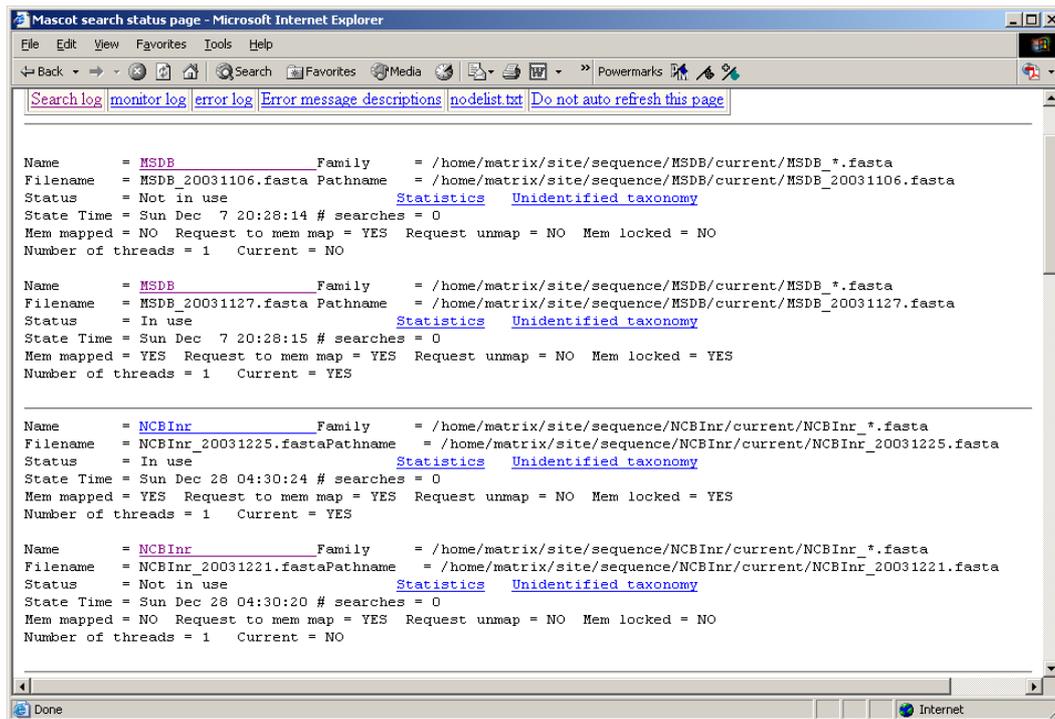
If the keyword `len` is supplied, then the length of the sequence is returned as `ascii` text. If the database is a nucleic acid database, then the length returned will depend on the translation frame number specified.

If the keyword `title` is supplied, the FASTA title line is returned, beginning with a right angle bracket.

If the keyword `pI` is supplied, the calculated iso-electric point is returned.

Status

The Status utility, `x-cgi/ms-status.exe`, provides an overview of the active and recent searches on all of the configured databases. The top level display will resemble this:



By clicking on a database hypertext link, a page is displayed showing the activity on that particular database:

Mascot database status - NCBIInr

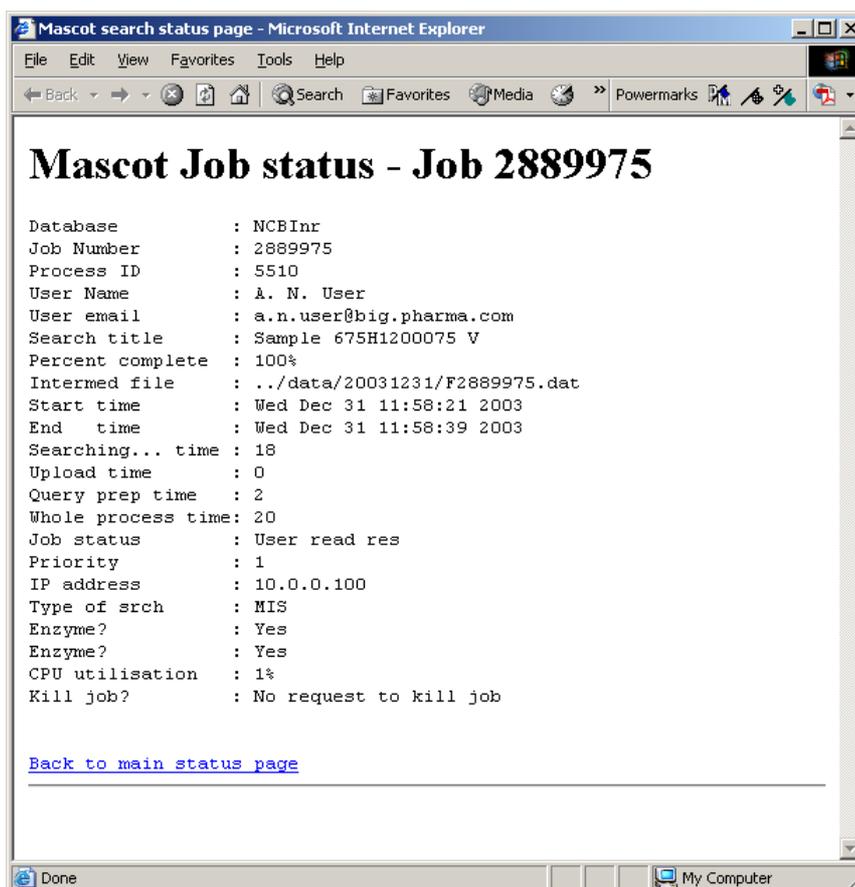
Current jobs

Job	PID	Start Time	Dur.	Status	User
2889975	16973	Wed Dec 31 11:58:21	2	Searching....	A. N. User

Completed jobs

Job	PID	Start Time	Dur.	Status	User
2889974	16973	Wed Dec 31 11:57:09	16	User read res	A. N. User
2889973	19928	Wed Dec 31 11:56:21	5	User read res	Pipeline 1
2889972	4772	Wed Dec 31 11:56:13	5	User read res	Ed Man
2889971	13815	Wed Dec 31 11:55:26	6	User read res	Pipeline 3
2889970	12756	Wed Dec 31 11:55:15	13	User read res	guest
2889969	30116	Wed Dec 31 11:55:06	6	User read res	Pipeline 1
2889968	26450	Wed Dec 31 11:54:37	17	User read res	A. N. User
2889967	17657	Wed Dec 31 11:53:37	15	User read res	A. N. User
2889966	9281	Wed Dec 31 11:52:59	5	User read res	Ed Man
2889965	16434	Wed Dec 31 11:52:03	16	User read res	A. N. User
2889964	32239	Wed Dec 31 11:50:13	12	User read res	guest
2889963	17772	Wed Dec 31 11:49:15	18	User read res	A. N. User
2889962	2322	Wed Dec 31 11:47:54	5	User read res	Pipeline 3
2889961	2695	Wed Dec 31 11:47:04	18	User read res	A. N. User
2889960	27572	Wed Dec 31 11:45:13	17	User read res	A. N. User
2889958	23741	Wed Dec 31 11:44:26	13	User read res	guest
2889955	8624	Wed Dec 31 11:43:14	13	User read res	Tevie

From which, links allow details of any specific search to be displayed:



Status can also be used to print Mascot configuration and result files to STDOUT. This provides a method to display these files in a browser. For example:

`http://your_server/mascot/x-cgi/ms-status.exe?Show=MS_ENZYMES`

where the argument to Show determines the file to be displayed:

MS_ENZYMES	enzymes
MS_FRAGMENTATION_RULES	fragmentation_rules
MS_MASCOT_DAT	mascot.dat
MS_MASSES	masses

MS_MOD_FILE	mod_file
MS_QUANTITATIONXML	quantitation.xml
MS_SUBSTITUTIONS	substitutions
MS_TAXONOMY	taxonomy
MS_UNIMODXML	unimod.xml
MS_USERS	users.xml

The above files are all displayed as plain text, without any formatting. If `Show=RESULTFILE`, then a results file from any directory under `mascot/data` can be returned, with HTML formatting. For example:

```
http://your_server/mascot/x-cgi/ms-status.exe?  
Show=RESULTFILE&DateDir=20031231&ResJob=F006983.dat
```

For security reasons, the following characters are not allowed in the `DateDir` or `ResJob`: `~ / \ :`

Review

Mascot Review, `x-cgi/ms-review.exe`, provides similar functionality to Status, but takes its input from `searches.log`. The tabular display can be filtered and sorted to locate specific searches by title, user name, or any one of the following log fields:

1. Mascot job number.
2. Process ID
3. Sequence Database
4. User name
5. User email address
6. Search title
7. Results file path
8. Start time and date
9. Duration in seconds
10. Completion Status
11. Job priority
12. Type of search: PMF, SQ, or MIS
13. Enzyme: Either `yes` (if user selected an enzyme) or `no` (if user selected enzyme type None).
14. User IP address

At the top of each column is a checkbox and a radio button. Select the radio button to sort the display on that column. Uncheck the checkbox to hide that column.

Along the top of the screen are a series of controls:

The Sort/filter button updates the display to reflect changes in parameters.

If you have multiple log files, a specific file can be displayed by entering its path into the Log File text field.

Start can be used to page through a long listing in blocks of entries specified by the number in the following field. Setting start to -1 displays the list starting from the last entry in the log file rather than the first

Finally, there is a field to specify a path to the data files. The log file only contains a relative path. If the data files have been moved, possibly to an archive directory or CD-ROM, the path to the new location can be specified here so as to restore the validity of the relative path.

An example of the Status display, filtered to show Sequence Query searches on the NCBIInr database is shown below:

Licensed to: Matrix Science Internal Test, (3 processors).

Sort / filter Log File: Start at: (-1=end, 1=start) how many: 1625 in log, 16 after filters. Data dir: GETs?:

Job#	PID	dbase	Us	Em	Ti	In	start time	Dur	Status	Pr	Typ	Enzyme	I
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>									
<input type="text"/>	<input type="text"/>	dbEST	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1603	6296	dbEST	Da	dc	Te	..	Mon Jul 31 18:15:05 2000	479	User read res		PMF	Yes	1
1596	6211	dbEST	Da	dc	Te	..	Mon Jul 31 17:34:49 2000	618	User read res		PMF	Yes	1
1595	6196	dbEST	Da	dc	Te	..	Mon Jul 31 17:32:16 2000	764	User read res		PMF	Yes	1
1594	6194	dbEST	Da	dc	Te	..	Mon Jul 31 17:31:43 2000	616	User read res		PMF	Yes	1
1584	6117	dbEST	Da	dc	Te	..	Mon Jul 31 17:17:14 2000	332	User read res		PMF	Yes	1
1577	6079	dbEST	Da	dc		..	Mon Jul 31 17:03:49 2000	458	User read res		PMF	Yes	1
1576	6076	dbEST	Da	dc		..	Mon Jul 31 17:02:40 2000	484	User read res		PMF	Yes	1
1575	6061	dbEST	Da	dc		..	Mon Jul 31 16:58:13 2000	243	User read res		MIS	Yes	1
1574	6041	dbEST	Da	dc		..	Mon Jul 31 16:51:16 2000	414	User read res		SQ	Yes	1
1571	4958	dbEST	Da	dc		..	Mon Jul 31 16:26:21 2000	405	User read res		SQ	Yes	1
1548	4518	dbEST				..	Mon Jul 31 13:25:21 2000	351	Job killed		MIS	Yes	1

Document: Done

GetTaxonomy

GetTaxonomy is a utility for retrieving taxonomy details for an entry in a database configured for use by Mascot.

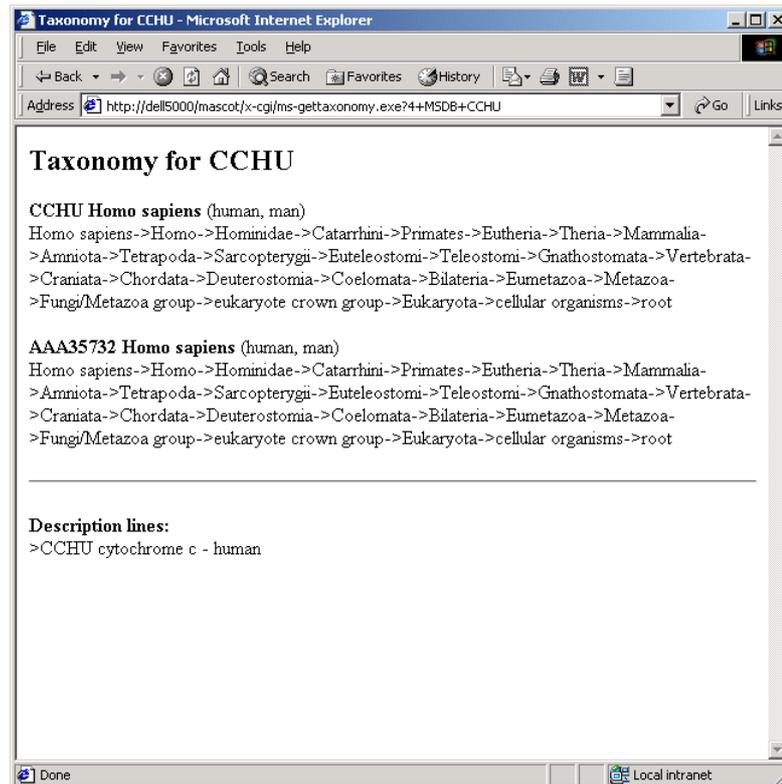
The executable, *x-cgi/ms-gettaxonomy.exe*, can be called from the command line, or as a URL.

When calling as a CGI application, with arguments appended to the URL, the parameter list must be URL escaped. (Spaces replaced by '+' and characters other than letters or numbers replaced by a '%xx' where xx is the ASCII code for the character as a hexadecimal number).

When running from a command line, the accession string should be enclosed in single or double quotes. This is essential for accession strings beginning `gi |`, because the pipe character has special meaning in Unix and NT. For example:

In the table below, the first argument supplied to *ms-gettaxonomy.exe* is an integer to specify the mode. The remaining arguments are selected from:

database	Mascot database name, e.g. MSDB
accession	accession string, e.g. CCHU
tax_ID	taxonomy ID number, e.g. 9606
species	name of species, e.g. "homo sapiens"



Parameters

1 database accession

2 database accession

Returns

Space separated list of accession string, tax_ID number, and scientific species name. Where a database entry represents multiple accessions, this information is repeated for each accession. Plain formatted.

Space separated pair of accession string and scientific species name. Where a database entry represents multiple accessions, this information is repeated for each accession. Followed by the FASTA title line for the accession supplied as an argument. Pretty formatted.

3 database accession	Same as mode 2, plus a list of common species names in parentheses.
4 database accession	Same as mode 3, plus complete taxonomy tree
5 database tax_ID	The scientific species name as a string. Pretty formatted.
6 database tax_ID	Same as mode 5, plus a list of common species names in parentheses.
7 database tax_ID	Same as mode 6, plus complete taxonomy tree
8 database species	verbose tax_ID information
9 database accession	genetic code number

SearchControl

Any helper application can call *bin/ms-searchcontrol.exe* to implement asynchronous automation of search submission. Available commands are:

```
--status  
--result_file_name  
--result_file_mime  
--result_file_ini  
--results  
--xmlresults  
--create_task_id  
--mascot_job_number  
--kill_job  
--pause_job  
--resume_job  
--nice_job  
--set_to_queued  
--version
```

ms-searchcontrol.exe --status --taskID <number> [--sessionID <string>]

The 'status' command will return one of the following:

unknown_id (referring to task ID)

id_assigned (referring to task ID)

error=nnnn

running=yy%

complete

queued

searchcontrol-error=nnn

where error indicates an error in the search, and will be the Mascot error number or one of:

TASK_ERROR_NO_ERROR	=	0
TASK_ERROR_JOB_CRASHED	=	-1
TASK_ERROR_JOB_KILLED	=	-2

And searchcontrol-error indicates a problem with the ms-searchcontrol.exe program. Values will be one of:

ERR_TASKID_NOERROR	=	0
ERR_TASKID_FAILOPEN	=	1
ERR_TASKID_FAILCREATE	=	2
ERR_TASKID_FAILREAD	=	3
ERR_TASKID_FAILWRITE	=	4
ERR_TASKID_FAILCLOSE	=	5
ERR_TASKID_CHANGEDRECORD	=	6
ERR_TASKID_INVALIDMASCOTDAT	=	7
ERR_TASKID_MISSINGRESULTSFILE	=	8
ERR_TASKID_FILENAMETOOLONG	=	9
ERR_TASKID_SESSIONTIMEDOUT	=	10
ERR_TASKID_PERMISSIONDENIED	=	11

ms-searchcontrol.exe --result_file_name --taskID <number> [--sessionID <string>]

This will return either the results file name:

filename=<filename>

or

```
searchcontrol-error=nnn
```

with values of 'nnn' as for --status.

Note that <filename> may be empty for some states – this is not an error.

This may then be used from the command line for other applications to provide functionality that is not in ms-searchcontrol.exe For example, a client application needs the USER name from a search. In this case, a perl script 'getusername.pl' could be written that takes the passed unique task ID, finds the results file name using:

```
ms-searchcontrol.exe --result_file_name
```

and then looks for the user name in the results file.

**ms-searchcontrol.exe --result_file_mime --taskID
<number> [--sessionID <string>]**

This will return the results file as a mime format file or

```
searchcontrol-error=nnn
```

with values of 'nnn' as for --status.

**ms-searchcontrol.exe --result_file_ini --taskID
<number> [--sessionID <string>]**

This will return the results file as a windows '.ini' format file or

```
searchcontrol-error=nnn
```

with values of 'nnn' as for --status.

**ms-searchcontrol.exe --results --taskID <number> [--
sessionID <string>]**

If the job is complete, then this will return the search results in a format recognised by Mascot Daemon:

For a peptide mass fingerprint, the output is of the form:

```
###daemon###file=..\data\F981122.dat  
###daemon###release=MSDB_20020121.fasta  
###daemon###queries=8  
###daemon###num_hits=6  
###daemon###h1=1A6K,103,1.00,17004.1  
###daemon###h1_text=myoglobin - sperm whale  
###daemon###reptype=concise  
###daemon###sigscoreprot=72  
###daemon###ionquery1=734.992175 from(736.000000,1+)
```

```

###daemon###ionquery2=746.992175 from(748.000000,1+)
###daemon###ionquery3=939.992175 from(941.000000,1+)
###daemon###ionquery4=1515.992175 from(1517.000000,1+)
###daemon###ionquery5=1591.992175 from(1593.000000,1+)
###daemon###ionquery6=1853.992175 from(1855.000000,1+)
###daemon###ionquery7=1980.992175 from(1982.000000,1+)
###daemon###ionquery8=2111.992175 from(2113.000000,1+)
###daemon###selectpeptides=0

```

For an MS/MS ions search, the output is of the form:

```

###daemon###file=..\data\F981123.dat
###daemon###release=MSDB_20020121.fasta
###daemon###queries=4
###daemon###num_hits=1
###daemon###h1=Q9XZJ2,286.477,1.00,79480.1
###daemon###h1_text=HEAT SHOCK PROTEIN 70.- Crassostrea gigas (Pacific
oyster) .
###daemon###reptype=peptide
###daemon###sigscoreprot=72
###daemon###ionquery1=1341.784350 from(671.900000,2+) query(1)
###daemon###score1=95.12
###daemon###sigscore1=49
###daemon###ionquery2=1614.584350 from(808.300000,2+) query(2)
###daemon###score2=74.55
###daemon###sigscore2=48
###daemon###ionquery3=1945.784350 from(973.900000,2+) query(3)
###daemon###score3=89.84
###daemon###sigscore3=47
###daemon###ionquery4=2167.784350 from(1084.900000,2+) query(4)
###daemon###score4=39.11
###daemon###sigscore4=47
###daemon###selectpeptides=1

```

If the job is incomplete, or has failed, then an error will be returned:

```

unknown_id
searchcontrol-error=nnn
with values of 'nnn' as for --status.

```

```

ms-searchcontrol,exe --xmlresults --taskID <number>
--reporttop [FILE|AUTO|num_hits] [--sessionID
<string>]

```

If the job is complete, then this will return the results formatted as an XML instance document that conforms to the schema

mascot/html/xmlns/schema/DistillerMascotSearch_1/DistillerMascotSearch_1.xsd

If the job is incomplete, or has failed, then an error will be returned:

unknown_id

searchcontrol-error=nnn

with values of 'nnn' as for --status.

ms-searchcontrol.exe --create_task_id [--sessionID <string>]

On failure, this will return

searchcontrol-error=nnn

with values of 'nnn' as for --status.

And on success it will return:

taskID=nnn

ms-searchcontrol.exe --mascot_job_number --taskID <number> [--sessionID <string>]

This will return either the job number:

mascotjobnumber=nnnn

or

searchcontrol-error=nnn

with values of 'nnn' as for --status.

ms-searchcontrol.exe --kill_job --taskID <number> [--sessionID <string>]

If the task is successful, this will return the text:

job_killed

If there is an error, one of the following will be returned:

unknown_id

job_not_running

searchcontrol-error=nnn

with values of 'nnn' as for --status.

The 'kill' is implemented by setting a flag in the mascot.control memory mapped file. The nph-mascot.exe task is responsible for 'killing' itself.

ms-searchcontrol.exe --pause_job --taskID <number> [--sessionID <string>]

If the task is successful, this will return the text:

job_paused

If there is an error, one of the following will be returned:

unknown_id

job_not_running

job_already_paused

searchcontrol-error=nnn

with values of 'nnn' as for --status.

The 'pause' is implemented by setting a flag in the mascot.control memory mapped file. The nph-mascot.exe task is responsible for 'pausing' itself.

ms-searchcontrol.exe --resume_job --taskID <number> [--sessionID <string>]

If the task is successful, this will return the text:

job_resumed

If there is an error, one of the following will be returned:

unknown_id

job_not_running

job_not_paused

searchcontrol-error=nnn

with values of 'nnn' as for --status.

The 'resume' is implemented by setting a flag in the mascot.control memory mapped file. The nph-mascot.exe task is responsible for 'resuming' itself.

ms-searchcontrol.exe --nice_job --taskID <number> [-nice <integer>] [--sessionID <string>]

The task ID need to be supplied. and an optional nice value

If a valid new nice value is supplied, this will return the text:

```
job_niced
```

If a nice value is not supplied, the program will return the current nice value:

```
nice=xxx
```

If there is an error, one of the following will be returned:

```
unknown_id
```

```
job_not_running
```

```
searchcontrol-error=nnn
```

with values of 'nnn' as for --status.

The 'nice' is implemented by setting a flag in the mascot.control memory mapped file. The nph-mascot.exe task is responsible for 'resuming' itself. Nice values range from -20 to +20. A value of +20 will set the task to a very low priority. The Mascot status screen shows the 'nice' value as a priority, which is simply $-1 * \text{the nice value}$. Microsoft Windows does not allow such a fine grained control of priorities, so, for example +20 and +19 will map to the same priority.

**ms-searchcontrol.exe --set_to_queued --taskID
<number> [--sessionID <string>]**

If the task is successful, this will return the text:

```
queued
```

If there is an error, one of the following will be returned:

```
unknown_id
```

```
already_running
```

```
already_complete
```

```
searchcontrol-error=nnn
```

with values of 'nnn' as for --status.

A batch processing client can make queued jobs visible to the Mascot system by getting a taskID and using this call to set the status to 'queued'. When the search is eventually submitted, nph-mascot.exe will set the status 'running'. A queued job will return 'queued' when ms-searchcontrol.exe is called with the --status argument.

```
ms-searchcontrol.exe --version [--sessionID  
<string>]
```

If the task is successful, this will return the version number

Miscellaneous Utilities

Service

Supplied for Windows only. This application shows the status of the Matrix Science Mascot Service, and allows it to be stopped and started. It is normally accessed from the start menu -

```
Programs; Mascot; config; Show Mascot Service Status
```

```
Programs; Mascot; config; Start Mascot Service
```

```
Programs; Mascot; config; Stop Mascot Service
```

These options run the program *x-cgi/ms-service.exe* with the first parameter set to the service name (`MatrixScienceMascotService`) and the second parameter being 0, 1, or 2 respectively.

It is also possible to run this program as a CGI script by entering the following URL in the browser:

```
http://your.host/mascot/x-cgi/ ↵  
⌘ ms-service.exe?MatrixScienceMascotService+0
```

Where *your.host* is replaced by the host name of the Mascot server. This CGI script can be run from any computer on the network. However, it is not usually possible to start and stop the service from another computer using the default access rights.

There is a final option, which will allow removal of the service. This may be required for a manual de-installation and will not normally be required. If this option is used, Mascot will not run again without re-running the installation program. The command to enter is:

```
ms-service MatrixScienceMascotService remove
```

Compress

Compress is a utility for compressing FASTA files independently of Mascot monitor.

The executable, *bin/ms-compress.exe* is executed from a shell or command prompt.

```
ms-compress.exe db_name fasta
```

where

`db_name` is the database family name from `mascot.dat` - e.g. MSDB

`fasta` is the fully qualified path to the FASTA file

`ms-compress.exe` compresses the fasta file using the rules specified in `mascot.dat` and must be run so that its current working directory is `mascot/bin`.

Return value of 0 for success, > 0 for failure

MakeSearchLog

The `bin/ms-makesearchlog.exe` utility is used to rebuild the search log by scanning all the result files located in sub-directories of `mascot/data`. This can be useful if the original search log has been damaged or if result files have been pruned after archiving. There are no arguments.

LockMem

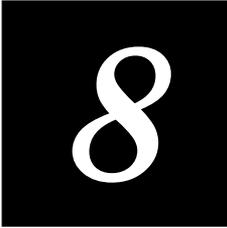
On 32 bit platforms, the 2Gb address space limit can quickly be reached by having several large databases locked into memory. To work around this limit, the `bin/ms-lockmem.exe` utility is provided.

LockMem is enabled by adding the parameter 'SeparateLockMem 1' to the options section of `mascot.dat`. Specifying a value greater than 1 specifies the block size in Mb. For example, if there is a 1.5 Gb *.s00 file, and this parameter is set to 750, two instances of `ms-lockmem.exe` will be run.

GetError

The utility `cgi/ms-geterror.exe` takes an error number as an argument and returns the corresponding text string. For example:

```
C:\Inetpub\MASCOT\cgi>ms-geterror.exe 23
You specified enzyme %s which is not available. Choose another.
```



I/O File Formats

Both Mascot search input files and results output files are in MIME format. This is a text file which can be viewed easily for inspection or debugging purposes.

The MIME format is defined in various “request for comment” documents. The following are the most relevant:

`ftp://ftp.isi.edu/in-notes/rfc2045.txt`

`ftp://ftp.isi.edu/in-notes/rfc2046.txt`

`ftp://ftp.isi.edu/in-notes/rfc2388.txt`

Very briefly, a unique boundary string is used to divide the file into sections, each of which contains data in a format defined by a Content-type.

Each section begins with two hyphens followed by the boundary string. The next line contains the content definition and name, followed by a blank line. Then data, until the beginning of the next section. For example:

```
MIME-Version: 1.0 (Generated by Mascot version 1.0)
Content-Type: multipart/mixed; boundary=gc0p4Jq0M2Yt08jU534c0p

--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="first_name"

first_value
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="another_name"

another_value
--gc0p4Jq0M2Yt08jU534c0p
.
.
.
--gc0p4Jq0M2Yt08jU534c0p
```

```
Content-Type: application/x-Mascot; name="final_name"
```

```
final_value  
--gc0p4Jq0M2Yt08jU534c0p--
```

Search Input File

The search input file is normally generated by the web browser. If another application is used to generate an input file, simply ensure that it conforms to the MIME format standard.

The Mascot Monitor test searches use “captured” input files. Hence, an example of a file can be seen by opening `mascot/data/test/SwissProt.asc` in any text editor.

```
-----16838575810113  
Content-Disposition: form-data; name="COM"
```

```
-----16838575810113  
Content-Disposition: form-data; name="USERNAME"
```

```
Monitor Program Test  
-----16838575810113  
Content-Disposition: form-data; name="DB"
```

```
SwissProt  
-----16838575810113  
Content-Disposition: form-data; name="SEARCH"
```

```
MIS  
-----16838575810113  
Content-Disposition: form-data; name="FORMAT"
```

```
Finnigan (.ASC)  
-----16838575810113  
Content-Disposition: form-data; name="FORMVER"
```

```
1.01  
-----16838575810113  
Content-Disposition: form-data; name="OVERVIEW"
```

```
OFF  
-----16838575810113  
Content-Disposition: form-data; name="REPORT"
```

```

AUTO
-----16838575810113
Content-Disposition: form-data; name="CLE"

Trypsin
-----16838575810113
Content-Disposition: form-data; name="MASS"

Monoisotopic
-----16838575810113
Content-Disposition: form-data; name="MODS"

SMA (K)
-----16838575810113
Content-Disposition: form-data; name="MODS"

SMA (N-Term)
-----16838575810113
Content-Disposition: form-data; name="IT_MODS"

Oxidation (M)
-----16838575810113
Content-Disposition: form-data; name="FILE"; filename="dyckall.asc"

LIST: dp210198b                21-Jan-98 DERIVED SPECTRUM    #9
Samp: Dyck tryptic morpholine fr 13      Start : 18:37:54    100
Mode: ESI +DAU 808.3 @ 25eV UP LR
Oper: Administrator                Inlet :
.
.
.
  734      1979.6      6      0.07      0.01      #
  735      1980.3      1      0.01      0.00      #
  736      1987.3      3      0.04      0.01      #
-----16838575810113
Content-Disposition: form-data; name="SEG"

-----16838575810113
Content-Disposition: form-data; name="CHARGE"

2+
-----16838575810113
Content-Disposition: form-data; name="PFA"

```

```
-----16838575810113
Content-Disposition: form-data; name="TOL"

0.1
-----16838575810113
Content-Disposition: form-data; name="TOLU"

%
-----16838575810113
Content-Disposition: form-data; name="LTOL"

50.0
-----16838575810113
Content-Disposition: form-data; name="ITOL"

0.5
-----16838575810113
Content-Disposition: form-data; name="ITOLU"

Da
-----16838575810113
Content-Disposition: form-data; name="QUE"

-----16838575810113-
```

Table 8.1 Mascot Search Parameters

Name	Description	Form type	Choices	PMF	SQ	MIS	Notes
ACCESSION	Database entries to be searched					y	Accession strings are quoted and comma separated
CHARGE	Peptide charge state	Select	'MH+', 'Mr', 'M-H-', '8-' to '8+' and combinations	y	y	y	Choice for PMF limited to 'M-H-', 'Mr', 'MH+'
CLE	Enzyme	Select	List from enzymes	y	y	y	Enzyme type None not available for PMF
COM	Search title	Text		y	y	y	
CUTOOUT	Precursor removal					y	See mascot.dat option PrecursorCutOut in Chapter 6
DB	Database	Select	List from mascot.dat	y	y	y	
DECOY	Flag for automatic decoy database search	Checkbox	0 (false), 1 (true)	y	y	y	
ERRORTOLERANT	Flag for error tolerant search	Checkbox	0 (false), 1 (true)			y	
FILE	Data File	File		y		y	
FORMAT	Data file format	Select	'Mascot generic', 'Sequest (.DTA)', 'Finnigan (.ASC)', 'Micromass (.PKL)', 'PerSeptive (.PKS)', 'Sciex API III', 'Bruker (.XML)', 'mzData (.XML)'			y	
FORMVER	Form version	Hidden					
FRAMES	NA translation		Comma separated list of frames				
INSTRUMENT	Instrument category	Select	List from fragmentation_rules		y	y	
INTERMEDIATE	Mascot result file containing ions data	Hidden					If there is a QueryX in the QUE field, load ions data from the Mascot result file specified here
IT_MODS	Variable modifications	Select	List from unimod.xml	y	y	y	An item can be selected in MODS or IT_MODS but not both.
ITOL	Fragment Mass Tol.	Text			y	y	
ITOLU	Units for ITOL	Select	'mmu', 'Da'		y	y	
MASS	Monoisotopic or Average	Radio	'Average', 'monoisotopic'	y	y	y	
MODS	Fixed modifications	Select	List from unimod.xml	y	y	y	An item can be selected in MODS or IT_MODS but not both
PEP_ISOTOPE_ERROR	Misassigned 13C	Select				y	
PFA	Partials Factor	Select	0 to 9	y	y	y	
PRECURSOR	Precursor mass	Text				y	For MS/MS file formats which do not specify precursor mass
QUANTITATION	Quantitation method	Select	List from quantitation.xml			y	
QUE	Query Window	Textarea		y	y		
REPORT	Number of hits to report	Select	AUTO, 5, 10, 30, 40, 50	y	y	y	List of values can be over-riden by ReportNumberChoices entry mascot.dat
REPTYPE	Type of report	Hidden	'Protein', 'Peptide', 'archive', 'concise', 'select', 'unassigned'	y	y	y	Default according to form type
RTINSECONDS	Retention time or range (in seconds)		a[[-b][,c[-d]]]			y	
SCANS	Scan number or range		v[[-w][,x[-y]]]			y	
SEARCH	Type of search	Hidden	'PMF', 'SQ', 'MIS'	y	y	y	Default according to form type
SEG	Protein Mass	Text		y	y	y	Units are kDa
TAXONOMY	Taxonomy	Select	List from taxonomy	y	y	y	
TITLE	Query title	N/A					Used in Mascot generic format file
TOL	Peptide Mass Tol.	Text		y	y	y	
TOLU	Units for TOL	Select	'mmu', 'Da', '%', 'ppm'	y	y	y	
USER??	User definable	Hidden		y	y	y	?? runs from 00 to 12
USEREMAIL	User email	Text		y	y	y	
USERNAME	User name	Text		y	y	y	

Results File

The results file contains the search results together with the search input parameters and MS data. This means that a results file contains everything necessary to generate a report, repeat the search at a later date, or act as the self-contained input file to a project database or LIMS.

The contents are divided into logical sections:

1. Search parameters
2. Mass values
3. Quantitation method (if used)
4. Unimod extract
5. Enzyme definition
6. Taxonomy (if a taxonomy filter was used)
7. Misc. header information
8. Summary results (for Protein Summary)
9. Mixtures (if PMF)
10. Summary of decoy results (if automatic decoy)
11. Summary of error tolerant results (if automatic ET)
12. Mixtures in decoy results (if automatic decoy PMF)
13. Peptides (if SQ or MIS)
14. Decoy peptides (if SQ or MIS and automatic ET)
15. Error tolerant peptides (if SQ or MIS and automatic ET)
16. Proteins (if SQ or MIS)
17. Query data, one block for each query
18. Index

General Notes

1. Values are shown in italics
2. Scripts are written so that label case doesn't matter.
3. Labels are used to assist readability, but kept short to minimise file size
4. Parameters are grouped logically
5. Order of blocks is not important except that the index block must be the last block. Presence of blank lines within the index block may cause a problem.
6. Because the MIME type is defined as an unknown application, if this file passes through a mail agent, it will be treated as an "octet stream" and encoded "base64" for transmission.

Search parameters

```
--gc0p4Jq0M2Yt08jU534c0p  
Content-Type: application/x-Mascot; name="parameters"
```

```
USERNAME=user name in plain text  
USEREMAIL=email address in plain text
```

```

SEARCH=PMF
COM=search title text
DB=MSDB
CLE=Trypsin
MASS=Monoisotopic
MODS=Mod 1,Mod 2
.
.
.
RULES=1,2,5,6,8,9,13,14
--gc0p4Jq0M2Yt08jU534c0p

```

The Parameters section contains the complete set of parameter values from the search form apart from the contents of the uploaded data file or the query window. Labels must be unique, independent of case. Where a parameter can be multivalued (e.g. mods) the values are listed on one line separated by commas.

RULES contains a list of the rule numbers that define the instrument type in the configuration file `fragmentation_rules`. The rule numbers are listed explicitly because the contents of the configuration file may have changed since the search was run.

Masses

```

--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="masses"

A=71.037110
B=114.534940
C=160.030649
D=115.026940
E=129.042590
F=147.068410
G=57.021460
H=137.058910
I=113.084060
J=0.000000
K=128.094960
L=113.084060
M=131.040480
N=114.042930
O=0.000000
P=97.052760
Q=128.058580
R=156.101110
S=87.032030

```

```
T=101.047680
U=150.953630
V=99.068410
W=186.079310
X=111.000000
Y=163.063330
Z=128.550590
Hydrogen=1.007825
Carbon=12.000000
Nitrogen=14.003074
Oxygen=15.994915
Electron=0.000549
C_term=17.002740
N_term=1.007825
delta1=15.994919,Oxidation (M)
NeutralLoss1=0.000000
FixedMod1=57.021469, Carbamidomethyl (C)
FixedModResidues1=C
--gc0p4Jq0M2Yt08jU534c0p
```

This block contains “actual” mass values. That is, average or monoisotopic residue masses, including any fixed modifications; C and N terminus groups also include any fixed modifications.

FixedMod1, FixedMod2, etc., records the delta mass and name for each fixed modification as comma separated values. FixedModResidues1 gives the site specificity. If multiple residues are affected, they are listed as a string, e.g. STY. If there was a neutral loss, the delta mass is given by the value of FixedModNeutralLoss1.

```
FixedModn=delta, Name
FixedModResiduesn=[A-Z]|C_term|N_term
FixedModNeutralLossn=mass
```

Fixed modifications cannot have peptide neutral losses, multiple neutral losses and cannot be protein-terminal or residue-terminal. In all these cases, fixed modifications are automatically converted into variable ones.

Variable modifications are reported in delta1, delta2, etc. Each entry defines the difference in mass introduced by the modification together with the name of the modification, separated by a comma. If a variable modification suffers a neutral loss on fragmentation, the delta is specified by a NeutralLossn entry. By definition, this is always a master neutral loss. If there are multiple neutral losses, then two more lines appear:

```
NeutralLossn_master=mass[[,mass] ...]
```

```
NeutralLossn_slave=mass[ [,mass] ...]
```

The first neutral loss (defined by *NeutralLossn*) has an implicit index number of 1. Any additional neutral losses (defined by *NeutralLossn_master* or followed by *NeutralLossn_slave*) have implicit index numbers of 2 and up.

If a modification includes a required or optional neutral loss from the precursor, this is recorded as follows:

```
ReqPepNeutralLossn=mass[ [,mass] ...]
PepNeutralLossn=mass[ [,mass] ...]
```

Error-tolerant modifications are not listed in masses section.

Quantitation

```
-gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="quantitation"

<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<quantitation majorVersion="1" minorVersion="0" xmlns="http://
www.matrixscience.com/xmlns/schema/quantitation_1" xmlns:xsi=
"http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://
www.matrixscience.com/xmlns/schema/quantitation_1 qu
antitation_1.xsd">
  <method constrain_search="false" description="15N metabolic label-
ling" min_num_peptides="2" name="15N Metabolic [MD]" pro
t_score_type="mudpit" protein_ratio_type="weighted"
report_detail="true" require_bold_red="true" show_sub_sets="0.5"
sig_th
reshold_value="0.05">
  <component name="light">
    <isotope/>
  </component>
  <component name="heavy">
    <isotope>
      <old>N</old>
      <new>15N</new>
    </isotope>
  </component>
</quantitation>
```

This section is an extract from *quantitation.xml* containing the quantitation method specified for the search. For more details and a link to the schema, refer to the Mascot HTML help pages for quantitation.

Unimod

```
-gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="unimod"

<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<umod:unimod xmlns:umod="http://www.unimod.org/xmlns/schema/unimod_2"
majorVersion="2" minorVersion="0" xmlns:xsi="http://w
ww.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://
www.unimod.org/xmlns/schema/unimod_2 unimod_2.xsd">
  <umod:elements>
    <umod:elem avge_mass="1.00794" full_name="Hydrogen"
mono_mass="1.007825035" title="H"/>
    <umod:elem avge_mass="2.014101779" full_name="Deuterium"
mono_mass="2.014101779" title="2H"/>
    <umod:elem avge_mass="6.941" full_name="Lithium"
mono_mass="7.016003" title="Li"/>
    <umod:elem avge_mass="12.0107" full_name="Carbon" mono_mass="12"
title="C"/>
```

This section is an extract from unimod.xml containing data for the elements, amino_acids, and any modifications specified in the search form. For more details and a link to the schema, refer to the help pages at www.unimod.org

Enzyme

```
-gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="enzyme"

Title:Trypsin
Cleavage:KR
Restrict:P
Cterm
*
```

This section is simply an extract from the enzyme file. Syntax details can be found in Chapter 6

Taxonomy

```
-gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="taxonomy"
```

```
Title: . . . . . Homo sapiens (human)
Include: 9606
Exclude:
*
```

This section is simply an extract from the taxonomy file. Syntax details can be found in Chapter 9

Header

```
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="header"

sequences=number of sequences in DB
sequences_after_tax=number of sequences after taxonomy filter
residues=number of residues in DB
distribution=see below
exec_time=search time in seconds
date=timestamp (seconds since Jan 1st 1970)
time=time in hh:mm:ss
queries=number of queries, (>= 1)
max_hits=maximum number of hits to be listed
version=version information
fastafilename=full path to database fasta file
release=filename of actual database used - e.g. Owl_31.fasta
taskid=unique task identifier for searches submitted asynchronously
pmf_num_queries_used=number of mass values selected for PMF match
pmf_queries_used=comma separated list of selected query numbers
Warn0=
Warn1=
Warn2=
--gc0p4Jq0M2Yt08jU534c0p
```

The Header section contains general values, used in the master results page header paragraph.

Distribution is a comma separated list of values that represent a histogram of the complete protein score distribution. The first value is the number of entries with score 0, the second is the number of entries with score 1, and so on, up to the maximum score for the search. Scores are converted to integers by truncation. This distribution is only meaningful for a peptide mass fingerprint search.

If intensity values are supplied for a peptide mass fingerprint, Mascot iterates the experimental peaks to find the set that gives the best score. The number of values selected is reported in `pmf_num_queries_used` and the selected queries listed in `pmf_queries_used`.

Summary results

```
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="summary"

qmass1=Mr
qexp1=m/z for query 1,
       charge
qintensity1=intensity value for query1 (if available)
qmatch1=Total number of peptide mass matches for query1 in database
qplughole1=Threshold score for homologous peptide match (MIS only)
qmass2=...
qexp2=...
qintensity1=
qmatch2=...
qplughole2=...
.
.
.
qmassn=...
qexpn=...
qintensityn=
qmatchn=...
qplugholen=...
num_hits=number of hits in the summary block (<= max_hits)
h1=accession string,
    total protein score,
    obsolete,
    intact protein mass
h1_text=title text
h1_frame=frame number (between 1 and 6, for nucleic acid only)
h1_q1=missed cleavages, (-1 indicates no match)
      peptide Mr,
      delta,
      start,
      end,
      number of ions matched,
      peptide string,
      peaks used from Ions1,
      variable modifications string,
      ions score,
      multiplicity,
      ion series found,
      peaks used from Ions2,
      peaks used from Ions3,
      total area of matched peaks
h1_q1_et_mods=modification mass,
```

```

        neutral loss mass,
        modification description
h1_q1_et_mods_master=neutral loss mass[[,neutral loss mass] ... ]
h1_q1_et_mods_slave=neutral loss mass[[,neutral loss mass] ... ]
h1_q1_primary_nl=neutral loss string
h1_q1_na_diff=original NA sequence,
                modified NA sequence
h1_q1_tag=tagNum:startPos:endPos:seriesID,...
h1_q1_drange=startPos:endPos
h1_q1_terms=residue,residue
h1_q1_subst=pos1,ambig1,matched1 ... ,posn,ambign,matchedn
h1_q2=...
.
.
.
h1_qm=...
h2=...
.
.
.
hn_qm=...
--gc0p4Jq0M2Yt08jU534c0p

```

Where a parameter has multiple values, these are shown on separate lines for clarity. In the actual result file, all values for a parameter are on a single line and there are no spaces or tabs between values.

Variable modifications is a string of digits, one digit for the N terminus, one for each residue and one for the C terminus. Each digit specifies the modification used to obtain the match: 0 indicates no modification, 1 indicates delta1, 2 indicates delta2 etc., in the masses section. If the number of modifications exceeds 9, the letters A to W are used to represent modifications 10 to 32. X is used to indicate a modification found in error tolerant mode.

neutral loss string is the same concept as the variable mod string, except each character represents the index of the primary neutral loss (one of the master NL). Any position that is not modified, or where the mod has no neutral loss, is set to 0. *hn_qm_primary_nl* will only be output if the string contains at least one non-zero character.

If a new modification is found in an error tolerant search, its position is marked by X, and details are recorded in an additional entry, *hn_qm_et_mods*. If the error tolerant search is of a nucleic acid database, and the modification is a single base change in the primary sequence, the two mass fields will be set to zero, and one of the keywords NA_INSERTION, NA_DELETION, or NA_SUBSTITUTION will appear in the

description field. The additional parameter `hn_qm_na_diff` is then used to record the 'before' and 'after' nucleic acid sequences.

Ion series is a string of 19 digits representing the ion series:

```
a
place holder
a++
b
place holder
b++
y
place holder
y++
c
c++
x
x++
z
z++
z+H
z+H++
z+2H
z+2H++
```

A digit is set to 1 if the corresponding series contains more than just random matches and 2 if the series contributes to the score.

Multiplicity means number of peptide mass matches for a query in a protein

For each sequence tag, four colon separated values are output: 1-based tag number, 1-based residue position where tag starts, 1-based residue position where tag ends, ion series into which the tag was matched:

```
-1 means no matches for the tag
0 "a" series (single charge)
1 "a-NH3" series (single charge)
2 "a" series (double charge)
3 "b" series (single charge)
4 "b-NH3" series (single charge)
5 "b" series (double charge)
6 "y" series (single charge)
7 "y-NH3" series (single charge)
8 "y" series (double charge)
9 "c" series (single charge)
10 "c" series (double charge)
11 "x" series (single charge)
12 "x" series (double charge)
13 "z" series (single charge)
```

```

14 "z" series (double charge)
15 "a-H2O" series (single charge)
16 "a-H2O" series (double charge)
17 "b-H2O" series (single charge)
18 "b-H2O" series (double charge)
19 "y-H2O" series (single charge)
20 "y-H2O" series (double charge)
21 "a-NH3" series (double charge)
22 "b-NH3" series (double charge)
23 "y-NH3" series (double charge)
25 "internal yb" series (single charge)
26 "internal ya" series (single charge)
27 "z+H" series (single charge)
28 "z+H" series (double charge)
29 high-energy "d" and "d'" series (single charge)
31 high-energy "v" series (single charge)
32 high-energy "w" and "w'" series (single charge)
33 "z+2H" series (single charge)
34 "z+2H" series (double charge)

```

If there are multiple tags for a query, comma separated groups of these numbers are output for each tag.

hn_qm_drang is output for a query that includes an error tolerant sequence tag. It defines the range of positions within which an unsuspected modification has been located. For a peptide of 10 residues, position 0 would indicate the amino terminus and position 11 would indicate the carboxy terminus. If there is no location information, the range is output as 0,256

hn_qm_terms shows the residues the bracket the peptide in the protein. If the peptide forms the terminus of the protein, then a hyphen is used instead.

hn_qm_subst is output when the matched peptide contained an ambiguous residue, (B, X, or Z). The argument is one or more triplets of comma separated values. For each triplet, the first value is the residue position, the second is the ambiguous residue, and the third is the residue that has been substituted to obtain the reported match.

For a large MS/MS search, *num_hits* is set to zero, and the summary block only contains entries for *qmassn*, *qexpn*, *qmatchn*, *qplugholen*. The threshold for switching to this mode is specified using two parameters in the Options section of *mascot.dat*. *SplitDataFileSize* is the size of the search process in bytes, (default 10000000), and *SplitNumberOfQueries* is the size of the search in queries, (default 1000).

If this is a two-pass search, either an automatic decoy database search or an automatic error tolerant search, a second summary block appears, containing the second set of results. The section name is either `et_summary` or `decoy_summary`. The syntax of the contents is identical

Mixture

```
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="mixture"

num_hits=number of mixtures found
h1_score=total score for mixture 1
h1_numprot=number of proteins in mixture 1
h1_nummatch=number of queries matched
h1_m1=accession string for protein component 1
h1_m2=accession string for protein component 2
.
.
.
h1_mm=accession string for protein component m
h2_score=
.
.
.
hn_mm=
--gc0p4Jq0M2Yt08jU534c0p
```

The Mixture section is only output for a peptide mass fingerprint. If any statistically significant protein mixtures are found, the mixture components are summarised. For details of individual components, use the accession strings to refer back to the Summary section.

If this is an automatic decoy database search, a second mixture block appears, containing the second set of results. The section name is `decoy_mixture`. The syntax of the contents is identical

Peptides

```
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="peptides"

q1_p1=missed cleavages, (-1 indicates no match)
      peptide Mr,
      delta,
      number of ions matched,
      peptide string,
      peaks used from Ions1,
      variable modifications string,
```

```

    ions score,
    ion series found,
    peaks used from Ions2,
    peaks used from Ions3;
    "accession string": data for first protein
        frame number:
        start:
        end:
        multiplicity,
    "accession string": data for second protein
        frame number:
        start:
        end:
        multiplicity,
    etc.
q1_p1_et_mods=modification mass,
    neutral loss mass,
    modification description
q1_p1_et_mods_master=neutral loss mass[[,neutral loss mass] ... ]
q1_p1_et_mods_slave=neutral loss mass[[,neutral loss mass] ... ]
q1_p1_primary_nl=neutral loss string
q1_p1_na_diff=original NA sequence,
    modified NA sequence
q1_p1_tag=tagNum:startPos:endPos:seriesID,...
q1_p1_drange=startPos:endPos
q1_p1_terms=residue,residue[:,residue,residue] ... ]
q1_p1_subst=pos1,ambig1,matched1 ... ,posn,ambign,matchedn
q1_p1_comp=quantitation component name
q1_p2=...
.
.
.
qn_pm=...
--gc0p4Jq0M2Yt08jU534c0p

```

Each line contains the data for a peptide match followed by data for at least one protein in which the peptide was found.

If there multiple entries in the database containing the matched peptide, there will be a corresponding number of pairs of bracketing residues listed in *qn_pm_terms*.

Otherwise, individual field descriptions are identical to those for the Summary section

If this is a two-pass search, either an automatic decoy database search or an automatic error tolerant search, a second peptides block appears,

containing the second set of results. The section name is either `et_peptides` or `decoy_peptides`. The syntax of the contents is identical

Proteins

```
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="proteins"

"accession string"=protein mass,
                    "title text"
.
.
.
"accession string"=protein mass,
                    "title text"
--gc0p4Jq0M2Yt08jU534c0p
```

This block contains reference data for the proteins listed in the peptides block.

Input data for query n

```
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="queryn"

Title=query title
Seq1=sequence qualifier (e.g. N-ABCDEF)
Seq2=...
.
.
.
Seqn=
Comp1=composition qualifier (e.g. 0[P]2[W])
Comp2=...
.
.
.
Compn=...
PepTol=peptide tolerance qualifier (e.g. 2.000000, Da)
IT_MODS=Mod 1[, Mod 2[, ...]]
INSTRUMENT=instrument identifier, (e.g. ESI-TRAP)
RULES=1, 2, 5, 6, 8, 9, 13, 14
INTERNALS=min mass, max mass
CHARGE=charge state (e.g. 2+)
RTINSECONDS=a[[-b][[, c[-d]]]]
SCANS=a[[-b][[, c[-d]]]]
Tag1=sequence tag (e.g. t, 889.4, [QK]S, 1104.54)
```

```

.
.
.
Tagn=...
mass_min=lowest mass
mass_max=highest mass
int_min=lowest intensity
int_max=highest intensity
num_vals=number of mass values
num_used1=-1 (obsolete)
ions1=1344.65:34.3,1365.41:13.2
ions2=y-1344.65:34.3,1365.41:13.2
ions3=b-1344.65:34.3,1365.41:13.2

--gc0p4Jq0M2Yt08jU534c0p

```

Value “query n ” runs from “query1” (no leading zeros). ions n values are sorted so that the matched values come first.

Most searches will only require a few of these fields. For example, a peptide mass fingerprint would only include the title field.

ions2 and ions3 are only required when fragment ions are specified in a sequence query as being N-terminal or C-terminal series.

The first field in a Tagn value is t for a standard sequence tag and e for an error tolerant sequence tag

Some search parameters can be define in the local scope of a query. These are CHARGE, COMP, INSTRUMENT, IT_MODS, TOL, TOLU. Any that are used are listed here. If the MGF file contained scan range information in terms of seconds or scans, this is written to RTINSECONDS and/or SCANS.

Index

```

--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="index"

parameters=4
masses=78
unimod=116
enzyme=322
taxonomy=329
header=336
summary=351
et_summary=6059
peptides=6473

```

```
et_peptides=7143
proteins=7292
query1=7362
query2=7374.
.
.
.
query81=8322
query82=8334
-gc0p4Jq0M2Yt08jU534c0p-
```

Values in index are the line number offsets of the section “Content-Type:” lines (starting from 0 for the first line of the file).

9

Taxonomy

Mascot supports the use of a species filter to limit the database entries to be searched. This is useful because it speeds up the search, and can reduce the number of hits in the results list to those more closely related to the sample being analysed.

Many public databases store the taxonomy in a manner that makes it difficult to extract species information reliably for all sequences. The major problems are:

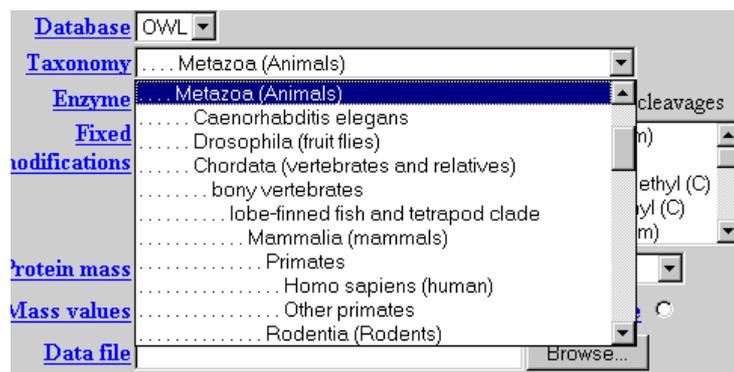
1. The location of the text containing the species identifier is mostly not defined, and can even vary within one database
2. There are often many names for one particular species - e.g. homo sapiens, human, man.
3. Names are sometimes misspelled - e.g. homo sapeins.
4. Continual re-classification of species is taking place
5. Some non-redundant databases only reliably give one species when several submissions from different species have identical sequences.
6. There are differences of opinion regarding the taxonomy 'tree' structure.

This section describes how the Mascot taxonomy filter works and how to configure it. Most of the configuration that will be required should be simple to change - for example the list of species displayed in the search form can be modified easily, and it is fairly simple to download updated taxonomy lists from the vendors of public web sites. However, to modify the configuration to use a new format and a different numbering system is a more complex task that may take some time.

For the supplied MSDB database, and also for the NCBI's nr, EST_others and SwissProt, it is recommended that the NCBI's taxonomy ID's are used. The NCBI keeps the list of ID's up to date, and guarantees that the ID for a given species will not change (although some of the names used for that ID may change). The configuration supplied uses the NCBI IDs, but it is possible to configure mascot to use a different system.

Modifying the list in the search form window

The list in the search form is taken from the *taxonomy* file in the *mascot config* directory.



This file can be edited using any text editor. (Under Windows, from the start menu, choose Programs, Mascot, Config, Mascot taxonomy file).

The following is an extract from the supplied file:

```
Title:All entries
Include: 1
Exclude: 0
*
Title:. . Archaea (Archaeobacteria)
Include: 2157
Exclude:
*
Title:. . Eukaryota (eucaryotes)
Include: 2759
Exclude:
*
Title:. . . Alveolata (alveolates)
Include: 33630
Exclude:
*
Title:. . . . Primates
Include: 9443
Exclude:
*
Title:. . . . . Homo sapiens (human)
Include: 9606
Exclude:
```

```

*
Title: . . . . . Other primates
Include: 9443
Exclude: 9606
*

```

The first line of each block must start with the `Title:` keyword, followed by a text string that is used to identify the species in forms and reports. The definition should be short and self-explanatory. To show the tree structure, indentation can be used. Unfortunately, it is not possible to use tabs or multiple spaces for indentation in an html form, so a full stop (period) and a space are used to indent the list. Internal spaces are significant, and there should never be two or more spaces together.

This should be followed with a definition line starting with the `In-clude:` keyword, followed by one or more numbers separated with commas. With the supplied MSDB database, and supplied taxonomy files, these numbers should be the NCBI taxonomy ID.

This should be followed with a definition line starting with the `Ex-clude:` keyword, followed by one or more numbers separated with commas. Any sequence with a taxonomy ID that passes the 'include' test, may then be rejected by any entry in the exclude list.

Finally, each entry must end with a *

There are two ways of finding the NCBI taxonomy ID for a given species. The first is to open the file `names.dmp` in the mascot taxonomy directory (Under Windows, from the start menu, choose Programs, Mascot, Config, NCBI taxonomy names.dmp file), and search for the species name. The ID is the number on the left. For example, the ID for Filicophyta is 3263:

3263	Filicophyta		scientific name	
3263	ferns		preferred common name	

Alternatively, the NCBI taxonomy browser can be used:

<http://www.ncbi.nlm.nih.gov/Taxonomy/>

Type the name into the query, and press the Start Search button. You may then need to click on the top where it says: "Click here to get information on this taxon."

For example, to add a choice of Ferns or human, add the following to the taxonomy file:

```

Title: . Ferns or human
Include: 9606, 3263
Exclude:
*

```

And to add the choice of Not human or mice add the following to the taxonomy file:

```
Title:. Not human or mice
Include: 1
Exclude: 9606, 10088
*
```

Note that 'all species', or root has the ID '1'.

It is, of course, possible to accidentally specify a selection that will result in no species matching - for example include humans, and exclude animals.

If you wish to include species in the taxonomy file without having them appear on the search form, the keyword 'Hidden' should appear on the line following the title line.

Location and format of species lists

NCBI Files

The NCBI provide two files that list all the species for which they have one or more sequences. These files are called *names.dmp* and *nodes.dmp*. As shown above, *names.dmp* is a list of scientific names, synonyms and misspellings for the species. From this list, you can easily find the ID for the given species. For example:

3701		Arabidopsis			scientific name	
3701		Cardaminopsis			synonym	
3702		Arabidopsis thaliana			scientific name	
3702		Arbisopsis thaliana			misspelling	
3702		thale cress			preferred common name	
3702		thale-cress			common name	
3702		mouse-ear cress			common name	

The file *nodes.dmp* specifies the tree structure. The first column is a taxonomy ID and the second column is the parent taxonomy ID. Note that the 'parent' of Arabidopsis thaliana (3702) is Arabidopsis (3701).

3700		3699		family		4		1		1		1		1		1		0		0		
3701		3700		genus		4		1		1		1		1		1		0		0		
3702		3701		species		AT		4		1		1		1		1		1		0		0

Both files can be obtained from the NCBI ftp site:

`ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz`

For NCBIInr, you will also need `gi_taxid_prot.dmp.gz`. For EST_others, you will also need `gi_taxid_nucl.dmp.gz`.

You should not modify the `names.dmp` and `nodes.dmp` file in the taxonomy directory. If you wish to add more entries, a new file should be made with just the new entries. Mascot will load multiple files as specified below. Most Mascot updates will contain the updated `names.dmp` and `nodes.dmp` files.

PDBeast File

This file contains a list of entries that are derived from the Brookhaven Protein databank (PDB). The file is available at:

`ftp://ftp.ncbi.nih.gov/mmdb/pdbeast/table`

SwissProt File

SwissProt also supplies a file, `speclist.txt` that is similar to the NCBI `names.dmp` file, except that it gives the NCBI taxonomy ID for the SwissProt Code:

Code	Taxon	N=Official name
	Node	C=Common name
		S=Synonym
AAV2	V 010804:	N=Adeno-associated virus 2 C=AAV2
ABDS2	B 056673:	N=Antarctic bacterium DS2-3R
ABIAL	E 045372:	N=Abies alba C=Edeltanne

This file is available at:

`ftp://ftp.ebi.ac.uk/pub/databases/uniprot/` 
 `knowledgebase/docs/speclist.txt`

If you wish to add more entries, a new file should be made with just the new entries- Mascot will load multiple files as specified below. Most Mascot updates will contain the updated `speclist.txt` file.

Genetic code selection

During a search of a nucleic acid database, Mascot also uses the taxonomy of each entry to choose the correct genetic code for translation. The genetic codes are defined in the NCBI file `gencode.dmp`, which is included in the archive `taxdump.tar`, mentioned above.

Nodes.dmp is used as a lookup table to obtain a genetic code number from a TaxID.

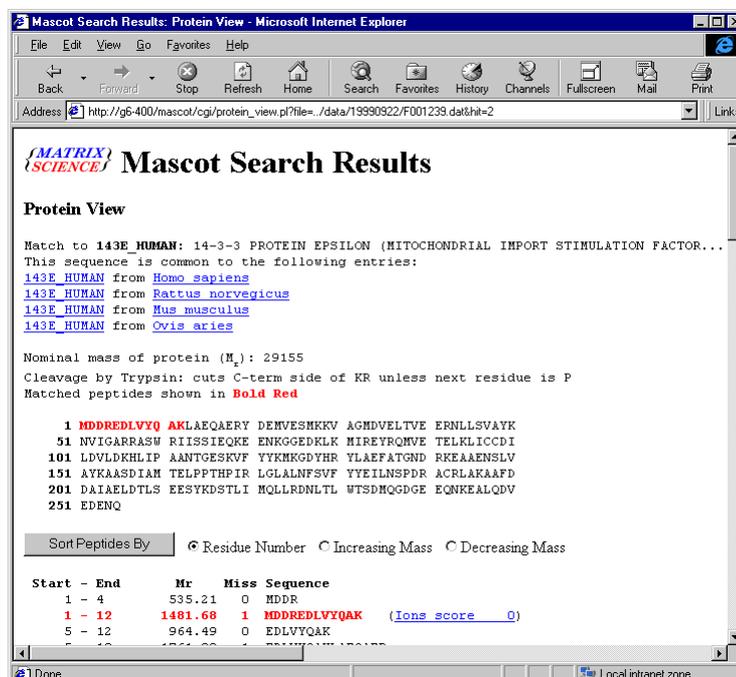
For many species, the genetic code is different for mitochondrial and nuclear proteins. Although Mascot could try to determine whether a database entry is mitochondrial by performing a keyword search of the FASTA description, this is unreliable. In any case, mitochondrial proteins will usually represent only a very small fraction of the entries in any comprehensive database. The most important requirement is to use the correct code for a database that is specifically mitochondrial proteins. The solution is to include a flag in each *mascot.dat* taxonomy definition to specify, at a database level, which code is to be used.

For further information on genetic codes, see:

<http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c>

Modifying the “Taxonomy lineage” link

In the protein view, a link to taxonomy lineage is shown:



The default behaviour is for this to link to the NCBI taxonomy browser. For non-redundant databases with more than one species source per

sequence, there will be a list of the species, each with a link. For the NCBI nr database, a separate 'gi' number will be shown for each database entry, with a link to Entrez and the NCBI Taxonomy browser for each entry.

If security and confidentiality protocols may make this unacceptable for your installation, then change the entry in the Options section of the *mascot.dat* file from:

```
TaxBrowserUrl http://www.ncbi.nlm.nih.gov/htbin-
  post/Taxonomy/wgetorg?lvl=0&lin=f&id=#TAXID#
to
```

```
TaxBrowserUrl ../x-cgi/ms-gettaxonomy.exe?4+#DATABASE##ACCESSION#
```

In this case, the link will display the information in the following format (this example is for NCBI rather than MSDB):

Taxonomy for gi | 4501885

gi | 4501885 Unknown species

gi | 113270 *Homo sapiens* (human, man, HUMAN, 103D, 10GS, 11GS, 121P, 12CA, 12GS, 133L)
 Homo sapiens->Homo->Hominidae->Catarrhini->Primates->Eutheria->Theria->Mammalia->Amniota->Tetrapoda->lobe-finned fish and tetrapod clade->bony vertebrates->Gnathostomata->Vertebrata->Craniata->Chordata->Deuterostomia->Coelomata->Bilateria->Eumetazoa->Metazoa->Fungi/Metazoa group->eukaryote crown group->Eukaryota->root

.

gi | 3320892 *Trichosurus vulpecula* (Tricosurus vulpecular, Trichosurus vulpecular, common brush-tailed possum, TRIVU)
 Trichosurus vulpecula->Trichosurus->Phalangeridae->Diprotodontia->Metatheria->Theria->Mammalia->Amniota->Tetrapoda->lobe-finned fish and tetrapod clade->bony vertebrates->Gnathostomata->Vertebrata->Craniata->Chordata->Deuterostomia->Coelomata->Bilateria->Eumetazoa->Metazoa->Fungi/Metazoa group->eukaryote crown group->Eukaryota->root

Description lines:

```
>gi | 4501885 | ref | NP_001092.1 | pACTB | beta actin
gi | 113270 | sp | P02570 | ACTB_HUMAN ACTIN, CYTOPLASMIC 1 (BETA-ACTIN)
gi | 71618 | pir | | ATHUB actin beta - human
gi | 71619 | pir | | ATMSB actin beta - mouse
gi | 279669 | pir | | ATCHB actin beta - chicken
gi | 28252 | emb | CAA25099 | (X00351) beta-actin [Homo sapiens]
gi | 49866 | emb | CAA27307 | (X03672) beta-actin (aa 1-375) [Mus musculus]
gi | 55575 | emb | CAA24528 | (V01217) beta-actin [Rattus norvegicus]
gi | 177968 (M10277) cytoplasmic beta actin [Homo sapiens]
gi | 211237 (L08165) beta-actin [Gallus gallus]
gi | 2116655 | dbj | BAA20266 | (AB004047) beta-actin [Homo sapiens]
gi | 2182269 (U39357) beta actin [Ovis aries]
gi | 2661136 (AF035774) beta actin [Equus caballus]
gi | 3320892 (AF076190) beta-actin [Trichosurus vulpecula]
```

Other parameters are possible for `ms-gettaxonomy.exe` - see the reference section 'ms-gettaxonomy' in Chapter 7.

Common Questions

Why do I sometimes get results for a species I didn't specify?

Sometime, when specifying for example 'Human' species, the results may appear at first sight to be from for example a Mouse sample. The most common reason for this is that for a non-redundant database such as MSDB, exactly the same sequence has been found in many species. To check this, look at the protein view, where you should see the species that you selected.

What is the "unclassified" and "other" species?

The NCBI cannot always classify every sequence - either because no species information was supplied with the data or because it currently doesn't fit into any currently classification. There are about 1500 such sequences in the NCBI nr database.

"Other" species include plasmids, and artificial sequences.

How do I see which sequences Mascot couldn't assign a taxonomy ID?

In the status screen, click on the "Unidentified taxonomy" link. This will show sequences where one or more of the species names were not identified by Mascot.

Why do I get the message "Taxonomy 'xxx' ignored. No taxonomy indexes for this database"

Check the following:

In the `mascot.dat` file, is parameter 14 for the problem database a valid number and is there a taxonomy section in `mascot.dat` for that number?

When the compressed files are built, the taxonomy index has the name 'database_name.t00' If this file doesn't exist for the database, it may be necessary to stop Mascot Monitor, delete the 'database_name.stats' file, and restart Monitor.

How Mascot gets a species ID for sequences

This section contains complex configuration information. It is normally only necessary to read and understand this section when adding a new database of a different type.

When ms-monitor creates the compressed files, it also makes a file containing the taxonomy ID(s) for each sequence. To do this it needs to follow certain rules. These rules are defined in mascot.dat. The rule number for each database is specified as the 14th parameter in the databases section of the mascot.dat file. To help explain these rules, the following sections describe these rules for MSDB, SwissProt, NCBI_{nr} and EST_others.

All text searches and comparisons are case insensitive, except where stated. Taxonomy definition keywords in the mascot.dat file are also case insensitive.

Several taxonomy definition blocks are obsolete, and retained only for backwards compatibility. Only the current definitions are described below.

NCBI_{nr}

The non-redundant protein database from the NCBI may contain many 'entries' (gi numbers) in each title line. These entries are separated by a control 'A' - decimal character '1'. The definition block for NCBI is number 8, and this contains the following:

```
# TAXONOMY FOR NCBInr using GI2TAXID
Taxonomy_8
Enabled          1    # 0 to disable it
FromRefFile      0
ErrorLevel       0
DescriptionLineSep 1    # ctrl a - hex code '1'.
SpeciesFiles     GI2TAXID:gi_taxid_prot.dmp, NCBI:names.dmp
NodesFiles       NCBI:nodes.dmp
DefaultRule      GI2TAXID,          CHOP: "gi|[0-9]*\"
Identifier        NCBI nr FASTA using GI2TAXID
AccFromSpeciesLine "\ (gi|[0-9]*\"
end
#
```

To turn off taxonomy for NCBI_{nr}, set the enabled flag to 0

FromRefFile is set to 0 indicates that the taxonomy should be found in the .fasta file rather than in a reference file.

ErrorLevel is set to zero, to indicate the type of warnings or errors that are found when creating the taxonomy information. If this is set to 0, then an entry is put into the 'NoTaxonomyMatch.txt' file for every sequence where **no** taxonomy information is found. If it is set to 1, then an entry is put into the file 'NoTaxonomyMatch.txt' for every sequence that had **any** gi number in a sequence without a match. Since some sequences will have up to 200 gi numbers (sources), there is a reasonable

chance that some of these entries will not have species information, and this would cause the errors files to become very large.

As described earlier, each entry in the description line is separated by a CTRL-A, so *DescriptionLineSep* is set to 1, the ASCII character code for CTRL-A.

There are two SpeciesFiles - *gi_taxid_prot.dmp* and *names.dmp*, and one NodesFile - *nodes.dmp*. All three files are available from the NCBI Web site.

The method of finding the species is particularly simple for NCBI databases. The default rule is applied to the FASTA title line to extract the 'gi' number from the accession string. The species file *gi_taxid_prot.dmp* is a look-up table that returns an NCBI taxonomy ID number.

For example, the FASTA title line

```
>gi|2147497|pir||S73153 hypothetical protein 10 - red
    ␣ alga (Porphyra purpurea) chloroplast
```

returns a gi number 2147497. Looking this number up in *gi_taxid_prot.dmp* returns a taxonomy ID of 2787. Looking this number up in *names.dmp* returns the string

```
2787 | Porphyra purpurea |
    ␣ scientific name |
```

MSDB

The taxonomy definition for MSDB is block 7:

```
# TAXONOMY FOR MSDB_20000621 or later
Taxonomy_7
Identifier          MSDB REF
Enabled             1 # 0 to disable it
FromRefFile         1
ErrorLevel          0
DescriptionLineSep  44 # 44 is a comma
ConcatRefFileLines  0
SpeciesFiles        NCBI:names.dmp, NCBI:owl.dmp
NodesFiles          NCBI:nodes.dmp
DefaultRule         NCBI, CHOP:WP "\(.*\)" # Everything!
PrefixRemoves       C;Species and mitochondrion chloroplast plastid
endogenous
SuffixRemoves
QuickRefSearch      C;Species
DoThisRuleFirst     "C;Species .*:\(.*)" # Everything after
TNU23504:
AccFromSpeciesLine  "C;Species \([^:]*\) "
end
#
```

To turn off taxonomy for MSDB, set the enabled flag to 0

FromRefFile indicates that the taxonomy should be found in the REF file rather than the Fasta file.

The line that contains the species IDs has multiple IDs, separated by a comma, so *DescriptionLineSep* is set to 44, which is the ASCII character code for a comma.

There are two SpeciesFiles used - names.dmp is the file available from the NCBI Web site as described above. Unfortunately, there are names used in MSDB that are not in the list provided by NCBI. In particular, abbreviations such as “E.coli” instead of “Escherichia coli” are sometimes used. Owl.dmp is in the NCBI format and currently contains about 200 entries.

Just one NodesFile is specified - nodes.dmp from the NCBI Web site.

The DefaultRule describes how to find the species name in the line of text in the SwissProt DAT file. The string in quotes is a regular expression (see Appendix A for a description of basic regular expressions). This is the simplest of all regular expressions and simply returns the complete string of text. The key NCBI indicates that the NCBI files specified should be used to search against the names. The key CHOP:WP describes how other parts of text are removed. The ‘P’ indicates that all words in the *PrefixRemoves* keyword should be removed before trying to do a match. The ‘W’ indicates that if there is no match to the complete string, then one word at a time should be removed from the *end* of the text, and the search performed again.

The *PrefixRemoves* list is short, but very necessary. Before searching, any of these words (followed by a space) are removed from the text. The first species in the list will always start with *C;Species:* because this is how the line is found.

The *QuickRef* search keyword is used to speed up the compressing of the database. Rather than use the regular expression for each line in the .ref file, this text is used for a fast compare to the string *C;Species*. Other lines are ignored. Note that this is a case sensitive compare, so it must be specified using the correct upper and lower case letters.

DoThisRuleFirst is applied first, and the default rule would only be used if this failed. MSDB explicitly associates each species line with the accession string of the primary database entry. A further rule, *AccFromSpeciesLine*, is used to extract this accession string. For example:

```
>P1;CCRB
cytochrome c [validated] - rabbit
C;Species CCRB: Oryctolagus cuniculus (domestic rabbit)
C;Species CYC_RABIT: Oryctolagus cuniculus (Rabbit).
```

For this entry, the QuickRefSearch rule would match twice. On the first occasion, *DoThisRuleFirst* would extract 'Oryctolagus cuniculus (domestic rabbit)' while *AccFromSpeciesLine* would extract 'CCRB'.

There is no match in names.dmp for 'Oryctolagus cuniculus (domestic rabbit)', so one word at a time would be removed from the end because of the CHOP:W. When the string is reduced to 'Oryctolagus cuniculus', this finds a match to taxonomy ID 9986.

The same process occurs for the second species, returning the same taxonomy ID.

SwissProt

The rules for SwissProt are fairly simple. Block 3 is used when only the FASTA file is available

```
# TAXONOMY FOR SwissProt from the fasta file
Taxonomy_3
Identifier          Swiss-prot FASTA
Enabled             1   # 0 to disable it
FromRefFile         0
DescriptionLineSep  0   # ctrl a - hex code '1'. For multiple descrip-
tions per entry
SpeciesFiles        NCBI:names.dmp, SWISSPROT:speclist.txt
NodesFiles          NCBI:nodes.dmp
DefaultRule         SWISSPROT, CHOP: ">[^_]*_\[^\ ]*\\" # Anything
after _ before space
PrefixRemoves
SuffixRemoves
QuickRefSearch
end
#
```

There is just a single species per sequence, so the DescriptionLineSep is set to 0.

The speciesFiles are just those from SwissProt, and the nodesFiles are taken from NCBI as before.

There is only one database source, so the DefaultRule can be used. This takes everything after the first underscore to the next space. For example:

```
>104K_THEPA (P15711) 104 KD MICRONEME-RHOPTRY ANTIGEN.
```

Would find the text THEPA, which it would look up in speclist.txt.

If the SwissProt DAT file is available, then the correct taxonomy definition is block 5:

```
# TAXONOMY FOR SwissProt from the .dat file
Taxonomy_5
Identifier          Swiss-prot DAT
Enabled             1   # 0 to disable it
FromRefFile        1
ConcatRefFileLines 1
ErrorLevel         2
DescriptionLineSep 44   # 44 is a comma
SpeciesFiles       NCBI:names.dmp, SWISSPROT:speclist.txt
NodesFiles         NCBI:nodes.dmp
DefaultRule        NCBI, CHOP:WP "\ (.*\)"
PrefixRemoves     OS and
SuffixRemoves
QuickRefSearch     "OS  "
end
#
```

FromRefFile indicates that the taxonomy should be found in the .ref file rather than the .fasta file.

ConcatRefFileLines is set to 1 because there may be multiple ‘OS’ lines listing all the species for a given entry. Where there are multiple species, the names are separated by a comma, so *DescriptionLineSep* is set to 44 which is the ASCII character code for a comma.

The *speciesFiles* file used to convert a species name to a taxonomy ID is the SwissProt speclist.txt file. Names.dmp and nodes.dmp are the NCBI files mentioned earlier.

The *DefaultRule* describes how to find the species name in the line of text in the SwissProt DAT file. The string in quotes is a regular expression (see Appendix A for a description of basic regular expressions). This is the simplest of all regular expressions and simply returns the complete string of text. The key NCBI indicates that the NCBI files specified should be used to search against the names. The key CHOP:WP describes how other parts of text are removed. The ‘P’ indicates that all words in the *PrefixRemoves* keyword should be removed before trying to do a match. The ‘W’ indicates that if there is no match to the complete string, then one word at a time should be removed from the *end* of the text, and the search performed again.

The *PrefixRemoves* list is short, but very necessary. Before searching, any of these words (followed by a space) are removed from the text. The first species in the list will always start with *OS* because this is how the line is found.

The *QuickRef* search keyword is used to speed up the compressing of the database. Rather than use the regular expression for each line in the .ref file, this text is used for a fast compare to any string beginning *OS* followed by 3 spaces. Other lines are ignored. Note that this is a case sensitive compare, so it must be specified using the correct upper and lower case letters.

EST_others

Definition 9 for *EST_others* is very similar to that for *NCBIInr*:

```
# TAXONOMY FOR dbEST using GI2TAXID
Taxonomy_9
Enabled          1    # 0 to disable it
FromRefFile      0
ErrorLevel       0
DescriptionLineSep 1    # ctrl a - hex code `1'.
SpeciesFiles     GI2TAXID:gi_taxid_nucl.dmp, NCBI:names.dmp
NodesFiles       NCBI:nodes.dmp
DefaultRule      GI2TAXID,          CHOP: "gi|\[0-9]*\"
Identifier       dbEST FASTA using GI2TAXID
GencodeFiles     NCBI:gencode.dmp
AccFromSpeciesLine "\ (gi|\[0-9]*\"
MitochondrialTranslation 0
end
#
```

A different species file, *gi_taxid_nucl.dmp*, is used for nucleic acid sequences.

The file containing genetic code data is specified as an argument to *GencodeFiles*, while *MitochondrialTranslation* is set to 0, specifying that all entries should be translated using the genetic code for nuclear proteins.

Species specific nucleic acid databases

Even a species specific database, such as *EST_human*, requires taxonomy to be defined at the database level, so that the correct genetic code can be chosen.

For *EST_human*, the default taxonomy block in *mascot.dat* is:

```
# TAXONOMY FOR EST_human with TaxID
Taxonomy_10
Enabled          1    # 0 to disable it
SpeciesFiles    NCBI:names.dmp
NodesFiles      NCBI:nodes.dmp
Identifier       EST_human FASTA with TaxID
GencodeFiles    NCBI:gencode.dmp
MitochondrialTranslation 0
TaxID           9606
end
```

MitochondrialTranslation is set to 0, (off), and *TaxID* is set to 9606, specifying that all database entries are homo sapiens. So, genetic code 1, (standard), will be selected for all entries.

10

Mascot Daemon

Overview

Mascot Daemon is a client application that automates the submission of searches to a Mascot Server. Functionality includes:

1. Batch mode, in which an arbitrary group of files can be defined for searching, either immediately or at some pre-set time.
2. Real-time monitor mode, in which new files on a pre-defined path are searched as they are created.
3. Data dependent follow-up tasks. For example, automatically repeating an unsuccessful search at a later date or against a different sequence database.

Tasks

The functional unit of Mascot Daemon is a task. A task can be created or modified in the Task Editor. A task is defined by:

1. The data source (e.g. a file list or a file path)
2. How the data are to be searched (an associated set of search parameters)
3. When the searches are to take place
4. Any follow-up activities, such as conditional repeat searches.

Tasks can be in one of four states: running, paused, completed, or cancelled. A paused task can be resumed. A paused or completed task can be cancelled or deleted.

Data Files

Data files can be any of the peak list formats supported by Mascot. Other types of file, such as binary data, can be specified if an appropriate data import filter is available:

1. A wide range of native file formats can be processed using the Mascot Distiller library, (requires an additional licence).
2. If Applied Biosystems | MDS Sciex Analyst is installed on the same system as Mascot Daemon, Analyst WIFF data files can be processed using the mascot.dll “script”.
3. If Applied Biosystems Data Explorer is installed on the same system as Mascot Daemon, Voyager DAT files can be processed.
4. If a licensed copy of lcq_dta.exe is installed on the same system as Mascot Daemon, Thermo Finnigan Xcalibur / BioWorks RAW files can be imported.
5. A utility called TS2Mascot can be used to import peak lists from an Applied Biosystems 4000 series database
6. If a Waters MassLynx sample list is specified, then Mascot Daemon will look for the *.pkl files corresponding to the entries in the sample list. Information from any of the fields in the sample list can also be embedded in the Mascot search title.

Flexibility

Several Mascot Daemon clients can submit searches to a single Mascot Server, and can even share a common task database. If you have several mass spectrometers, you can choose whether to install separate copies of Daemon on each instrument data system or whether to have a single copy of Daemon somewhere on the LAN, marshalling searches for all instruments.

User Help

Mascot Daemon includes comprehensive, context sensitive on-line help. Press F1 at any time to jump to the relevant topic.

Installation

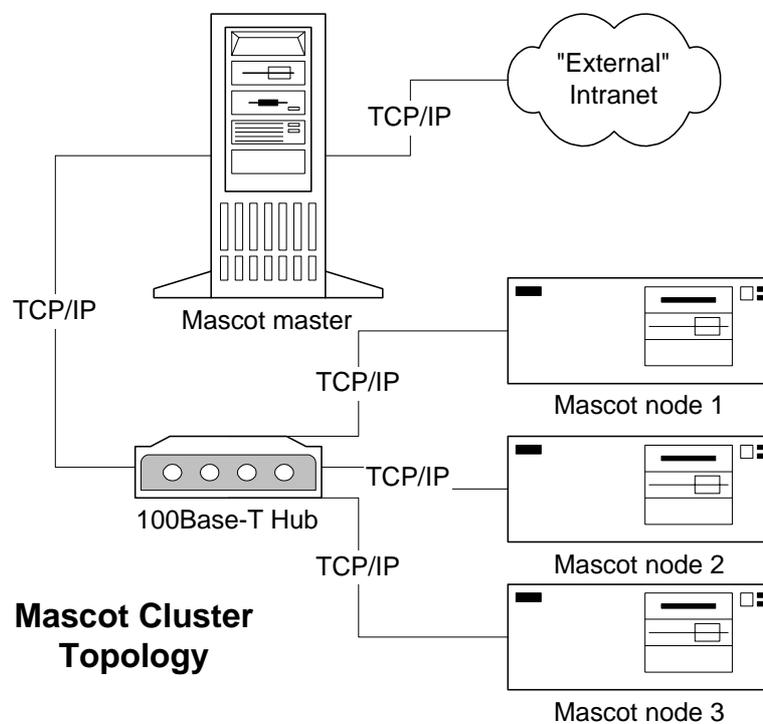
After Mascot Server has been installed, go to your local home page for links to a help page that describes how to install, upgrade or troubleshoot Mascot Daemon. All the required installation files are hyperlinked from this page.

11

Cluster Mode

Introduction

Mascot has been designed and implemented to work efficiently on a cluster of computers. A cluster of single or dual processor boxes provides a highly cost effective solution for high throughput protein identification. Mascot can be run in cluster mode on all supported hardware platforms and operating systems.



Hardware Requirements

All machines in a cluster should have processors of the same speed. Otherwise, the box with the slower processor(s) will become a bottleneck.

Two network ports are required on the master server; one for external access and one for communication on the private local area network (LAN) that connects the master to the nodes. The LAN for the cluster should run at least at 100Mb/s.

The total amount of RAM required in a cluster is a function of how many sequence databases need to be memory mapped simultaneously. Mascot supports up to 64 databases, but only those that are searched regularly or used for high throughput work need to be locked in memory. The others can be allowed to swap in and out of memory as needed. For example, a 5 node, 10 processor cluster might have 2 Gb on the master, and 1 Gb on each of the nodes. Allowing 100 Mb for the operating system on each node leaves 900 Mb per node or 4.5 Gb in total for searches and databases. This would be sufficient to have NCBI nr and all three EST databases locked in memory, yet still leave enough room for other databases to swap in and out of memory as needed.

Each node requires sufficient free disk space for the Mascot application software and the compressed FASTA databases. Since hard disks are very inexpensive, it makes sense to specify a hard disk of at least 75 Gb for each node. The master also requires sufficient disk space for the original FASTA databases and the accumulating search result files. The amount of space required for the results files depends on how heavily the system is used and how often the files are backed-up and deleted. A pair of 75 Gb hard disks might be a good start.

For best performance, it is advisable for the nodes to have local hard disks. If you prefer to use shared storage, then each node must have its own dedicated directory structure.

Mascot nodes may have any number of processors.

A video card is required in order to boot up Windows, but this can be the least expensive available since it will have no effect on performance

A node does not require a keyboard, monitor or mouse. If you are running Windows on the nodes, and want to be able to “see” the individual desktops, you might consider using a KVM switch so that a single keyboard, monitor and mouse can be shared between all the nodes. Alternatively, remote desktop viewing software, such as VNC, can be used.

<http://www.realvnc.com/>

Operating System Requirements

For nodes running Windows, it is not necessary to use a 'Server' version of Windows on the slaves. Windows 2000 Professional, XP Professional, or Vista are fine.

For Unix clusters, it must be possible for the master to communicate with the slave nodes using ssh or rsh without quoting a password or passphrase.

Slave nodes do not require Perl or web server software.

The master detects that slave nodes are responding by using ping under Unix and ICMP echo under Windows. Hence, these services must not be disabled or blocked by firewalls

Overview of Implementation

Each search is distributed to all the cluster nodes, but each node searches just an allocated portion of the sequence database. Search results are returned to the master, which merges them, writes the result file to disk, and optionally generates HTML reports to be returned to a client web browser.

All master - node communication is via TCP/IP.

Configuration and program files are distributed and updated automatically from the Master node.

Mascot administration tools provide web browser based system status reports. These are continuously updated and show at a glance important parameters such as processor usage and free disk space for each of the nodes. As an option, critical alerts can also be sent to the system administrator by email.

In cluster mode, Mascot is intended to run as a dedicated system. Trying to run other applications on the cluster simultaneously may have unpredictable effects on search speed.

Very large clusters

Very large clusters (> 30 nodes) pose certain special problems:

- Even with reliable hardware, node failures can be expected relatively frequently
- LAN communication can become a bottleneck

Mascot allows large clusters to be divided into sub-clusters. Each sub-cluster uses identical databases and configuration files, but operates

independently of the other sub-clusters. An incoming search is directed to the first available sub-cluster.

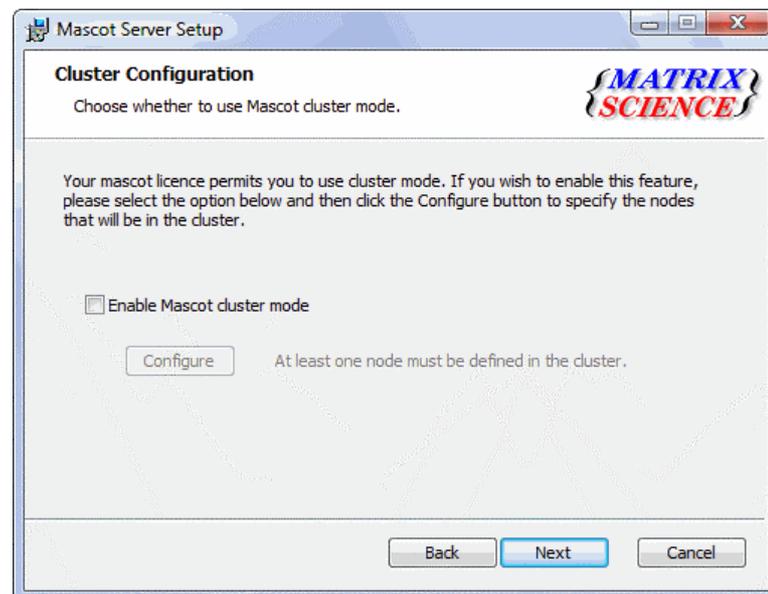
Should a node go down, only the sub-cluster is affected. Ideally, there will be one or more 'spare' nodes defined. Mascot will reconfigure the sub-cluster to include a spare node and re-start. If there are no spare nodes, Mascot will reconfigure the sub-cluster to exclude the faulty node and re-start.

Installation of Mascot

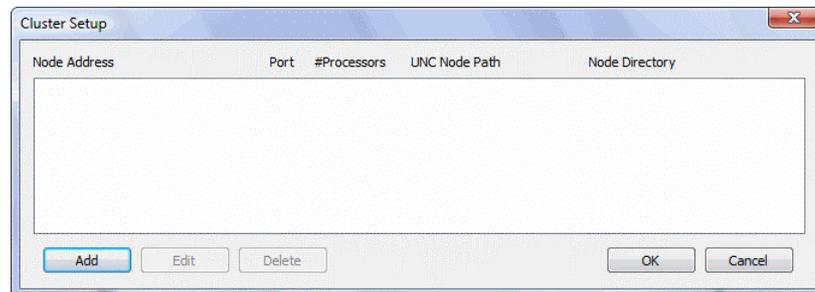
It is only necessary to install or upgrade Mascot on the master system. In fact, no files are copied to the Mascot nodes during installation. The distribution of files and executables is all handled when Mascot Monitor starts.

Windows

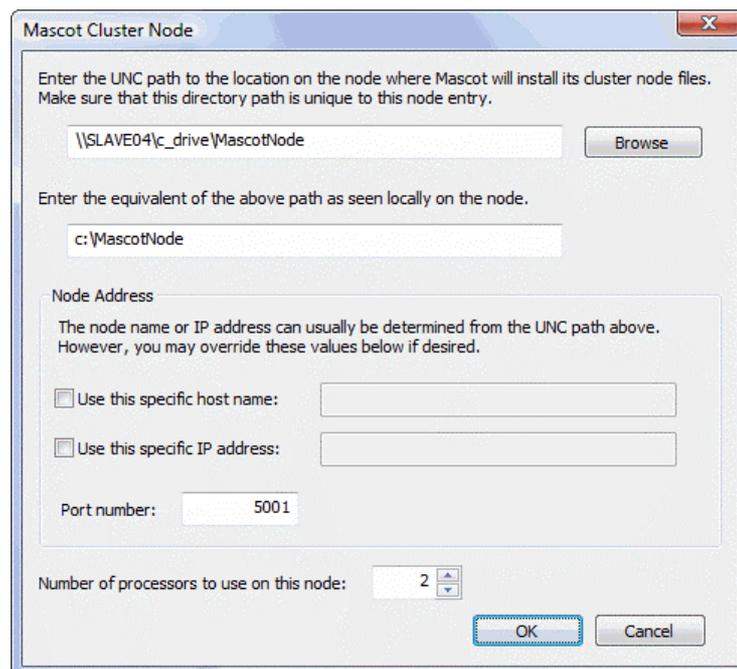
During Mascot installation on a Windows system, if your Mascot licence is for 4 or more processors and it is a first time installation, the following dialog will be displayed:



If you enable cluster mode, the configure button invokes the following dialog



Choose Add to configure each cluster node



Use the Browse button to ensure that the UNC path to the node is correct. If the machines are in a Windows domain, and the remote drive is not explicitly shared, you can enter C\$ for drive C, etc., to use the administrative shares. If the base directory does not exist, create it using the 'Make New Folder' button. The recommended base folder name is MascotNode.

Ensure that the local path to the MascotNode directory matches the UNC path. This must be a local or mapped drive on the node so that the path can be specified using a drive letter. The dialog will try to guess the

local path from the UNC path, but it may get it wrong. Ensure that this path is correct before pressing OK.

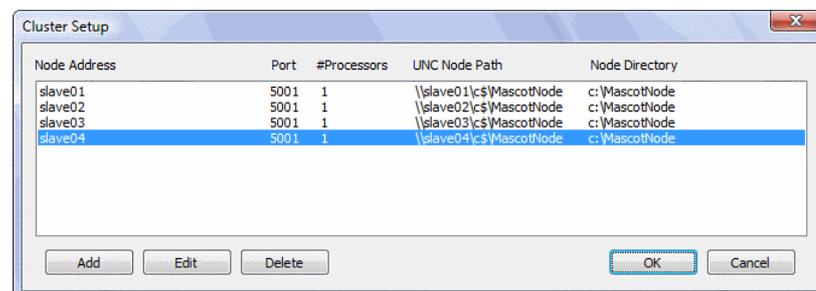
It is not necessary to fill in the Host name and IP Address fields unless the node is a multi-homed system and it is necessary to define which network interface will be used for communication with the Mascot master.

The default port number for cluster communication is 5001. If there are conflicts, this can be changed

The number of processors must be specified. The total number of processors specified for all nodes can be greater than the number of processors in the Mascot licence. The surplus processors will then behave as 'hot-spares', to be swapped into the cluster as required if there is a hardware problem on another node.

NB If the master is also a search node, and will execute Mascot searches in addition to running Mascot Monitor and the web server, it must be defined as a node using this dialog.

Use the Add, Edit, and Delete buttons to specify the complete cluster. After adding several nodes, the cluster setup dialog will resemble this.



Press OK to return to the installation wizard, and file installation will begin. Copying the files and configuring the system may take some time.

Once complete, you will be presented with a message requesting that you configure and start the Mascot Monitor service. This has to be done manually, because the Monitor service on the master needs to be run as Administrator so that it can write to the registry, install, start and stop services on the cluster nodes. (NB If you later change the password for Administrator, you must also change it in the Logon tab of the Matrix Science Mascot Service properties).

Very large clusters

Defining a very large cluster using the Add node... dialog can be tedious. It is usually faster to define a small cluster, let the installation program run to completion, then edit the configuration files using a text editor.

From the Program menu, stop the Mascot service, and edit the cluster and sub-cluster configuration details into *mascot.dat* and *odelist.txt*. A full description of these files can be found below in the 'Reference' section. Then, start the Mascot service.

Windows Firewall (Part 1)

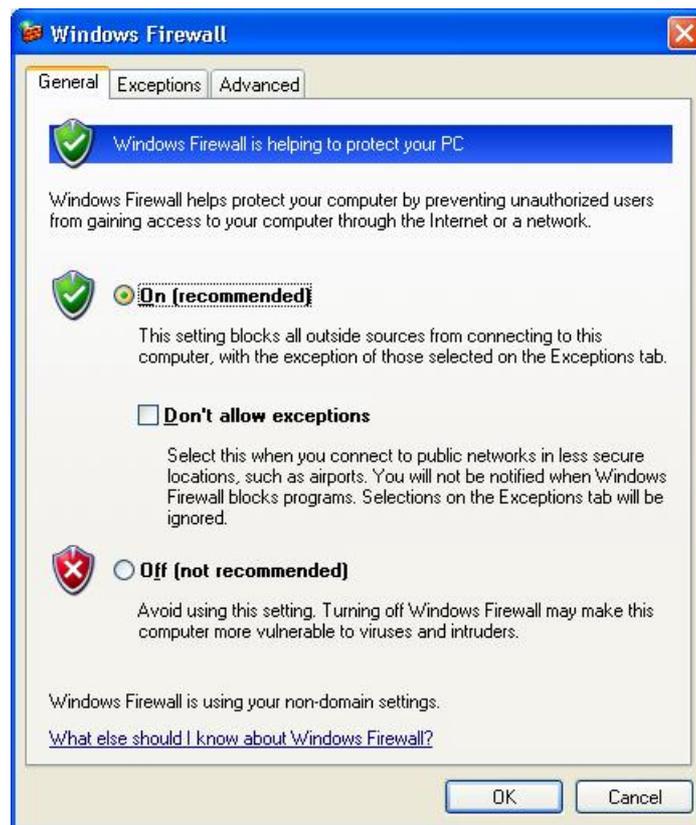
Windows firewall configuration varies across the different editions of Windows and also according to whether it was part of the original installation or added in a service pack. The following dialogs show typical dialogs for Windows XP SP2 and Server 2003 SP1. For Windows Vista, refer to the release notes.

You can choose to turn the Windows firewall off. If you prefer to have it on, as recommended by Microsoft, it is necessary to run through the following steps on each slave node and make changes as necessary

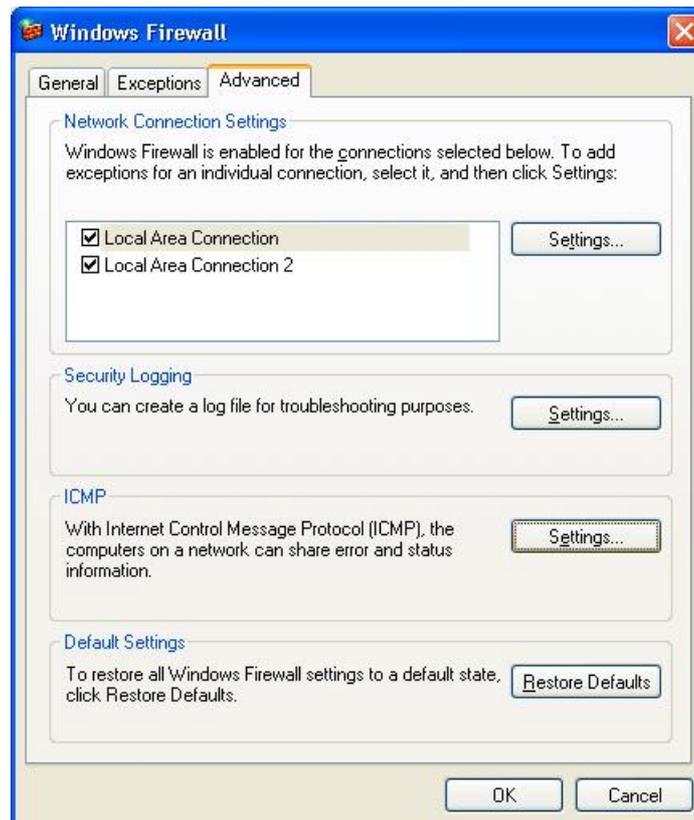
Log in as a local administrator. Go to Control Panel and launch Network Connections. Select the active network connection and right-click on it. Select Properties from the context menu. In Properties window select the tab Advanced as shown below.



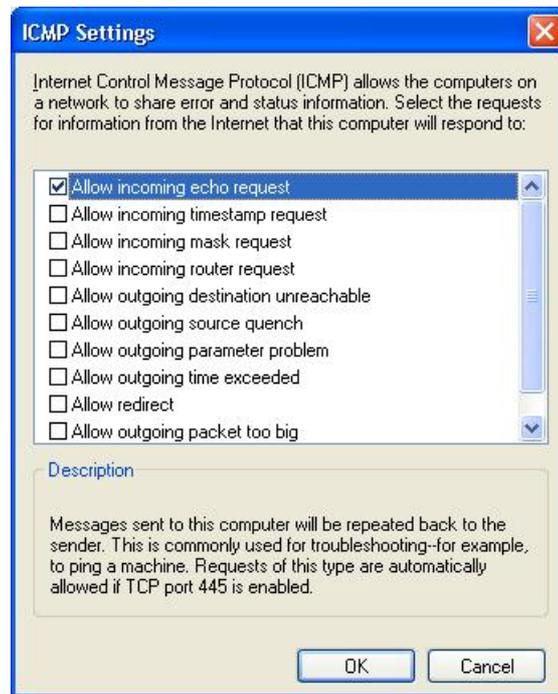
Choose the Windows Firewall Settings button to open the following dialog.



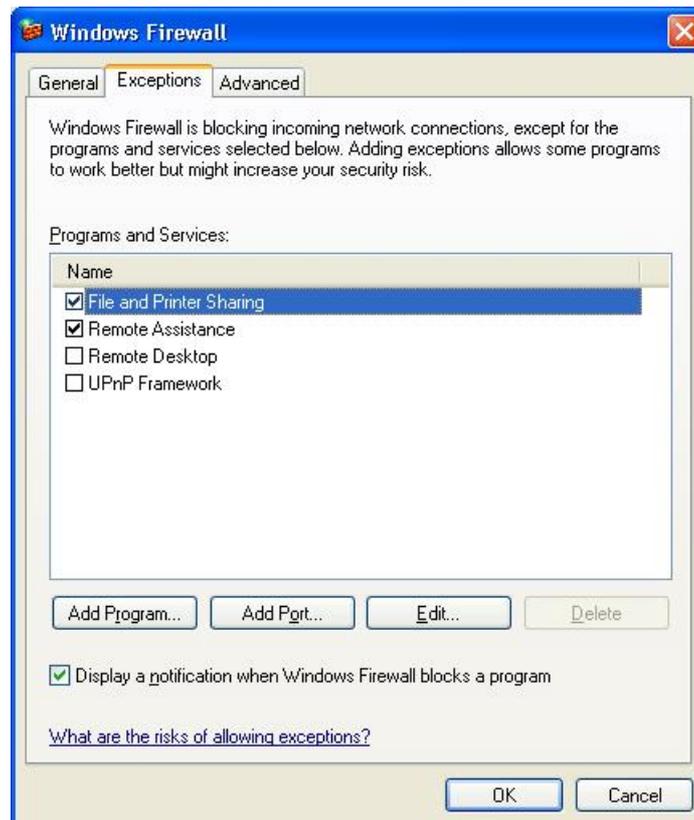
Microsoft recommends having the Windows firewall enabled. Go the Advanced tab and choose ICMP Settings



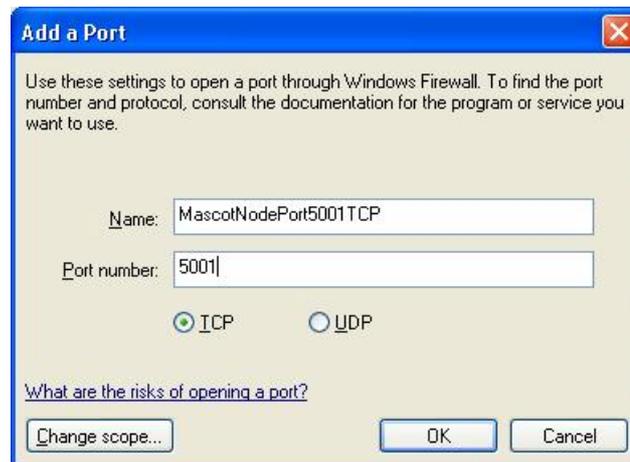
Check *Allow incoming echo request* and choose OK.



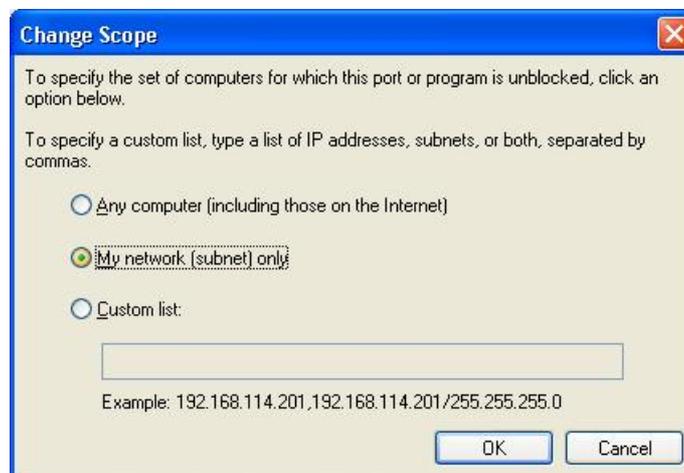
Now, go to the Exceptions tab and ensure that *File and Printer Sharing* is checked



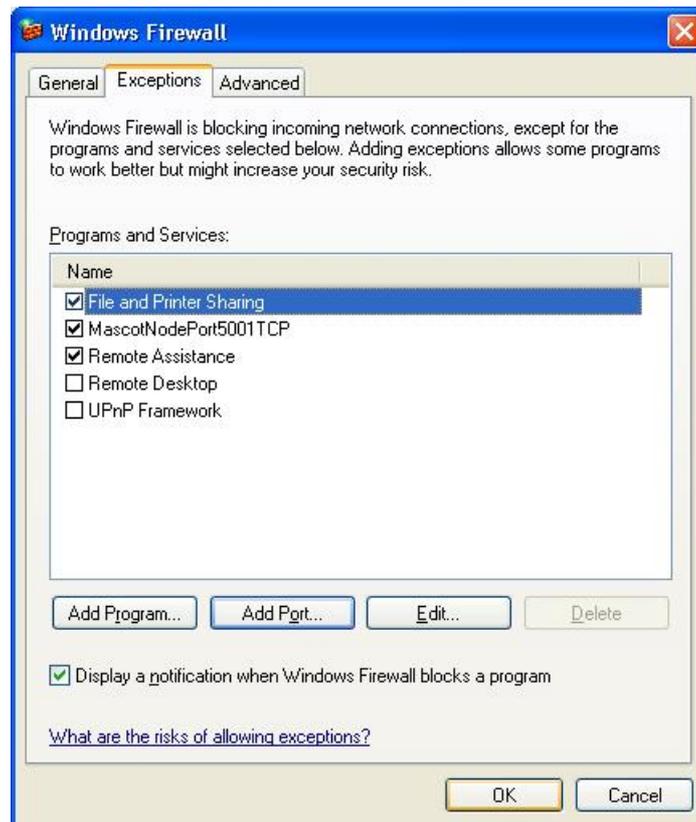
Choose *Add Port* and enter the following data. **Don't** press OK yet.



Choose *Change scope* and select the second option: *My network (subnet) only*.



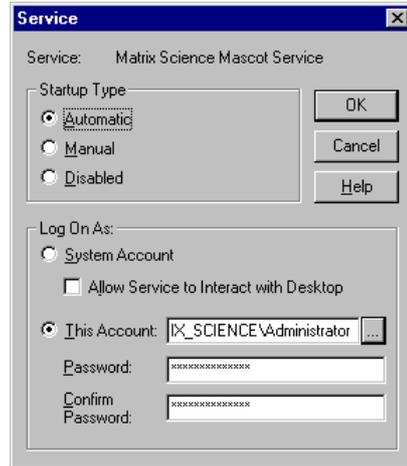
Now press OK buttons in this and in the previous window in order to return to the Windows Firewall dialog, which should now look like this



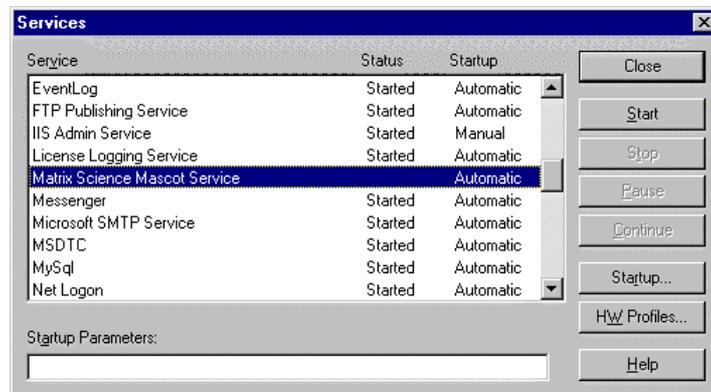
Press OK. Once the firewall on all the slave nodes has been configured, the Mascot service must be started before the remainder of the firewall configuration can be completed

Starting the Mascot service for the first time

From the Control panel, open the Services dialog and choose Matrix Science Mascot Service. Select Startup... and change the "Log On As" to the domain Administrator account. Note that, if the cluster nodes belong to a domain, it is essential to specify the domain Administrator account, and not the local (machine) Administrator account. If the nodes do not belong to a domain, either all nodes must share the same Administrator password or all the Administrator passwords must be empty.



Press OK, and you will be returned to the Services dialog:



Highlight the entry for Matrix Science Monitor Service and press Start.

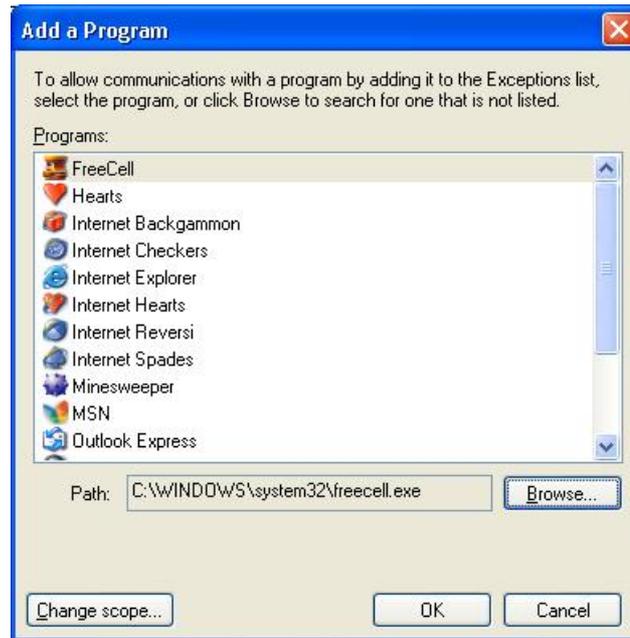
Windows Firewall (Part 2)

Wait until all the program files and database files have been copied to the slave nodes. You will observe that the cluster fails to reach the In Use state because some additional firewall configuration is required

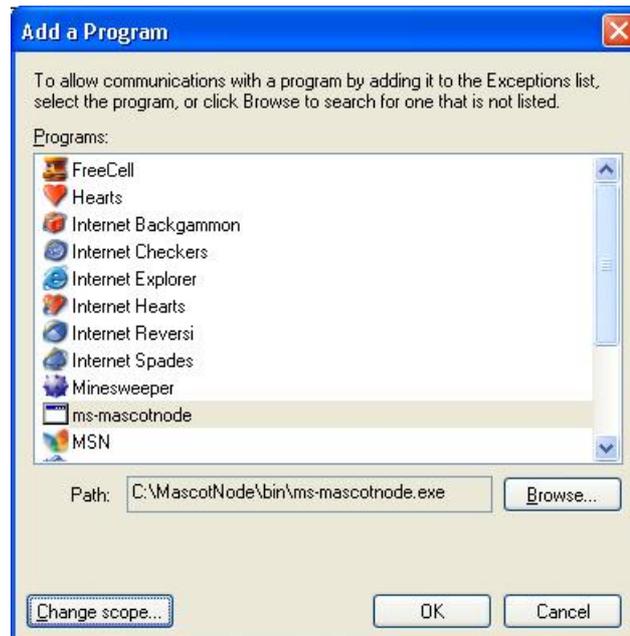
Stop Mascot service and run through the following steps on each slave node.

From the control panel, open Network Connections. Select the active network connection and right-click on it. Select Properties from the context menu. In the Properties dialog, select the Advanced tab. Choose

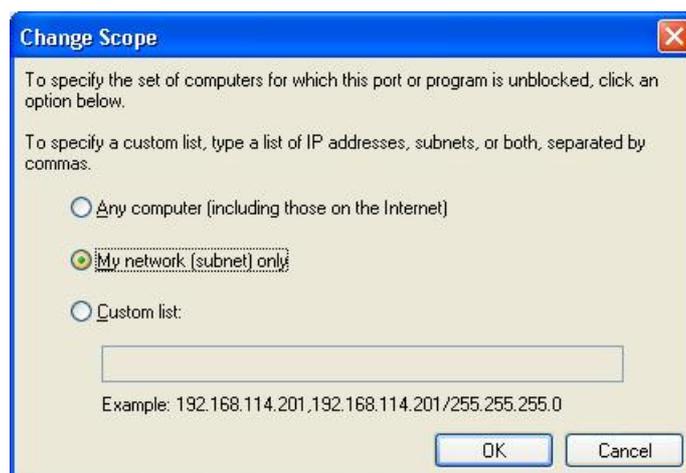
Windows Firewall Settings. Select the Exceptions tab and then *Add Program* to display this dialog



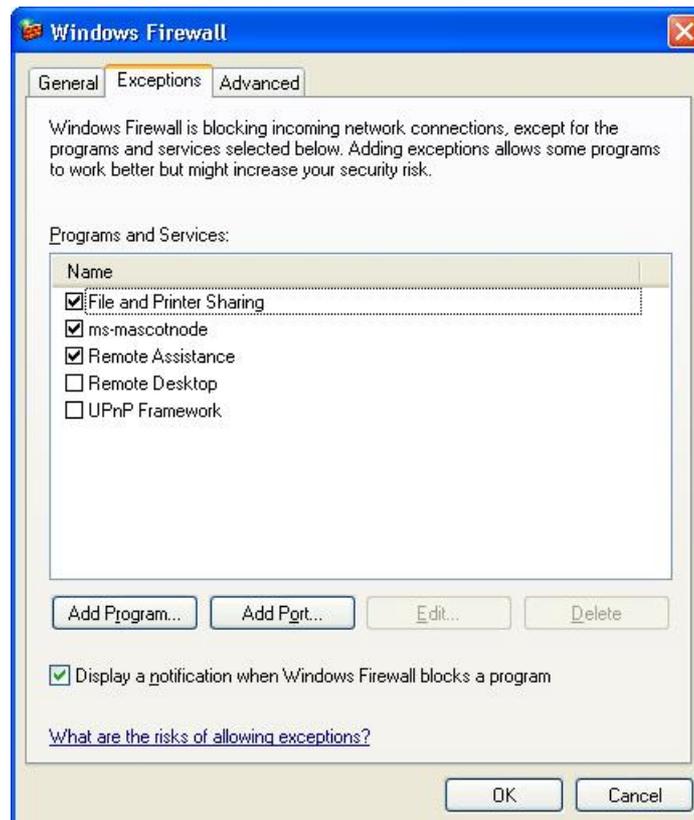
Click on Browse button and find ms-mascotnode.exe on the local computer, (probably in C:\MascotNode\bin). This program will be added to the display



Select ms-mascotnode and choose *Change scope*. Select *My network (subnet) only*.



Press OK twice to return to the Windows Firewall window. Select MascotPort5001TCP and press Delete button. The dialog should look now like this

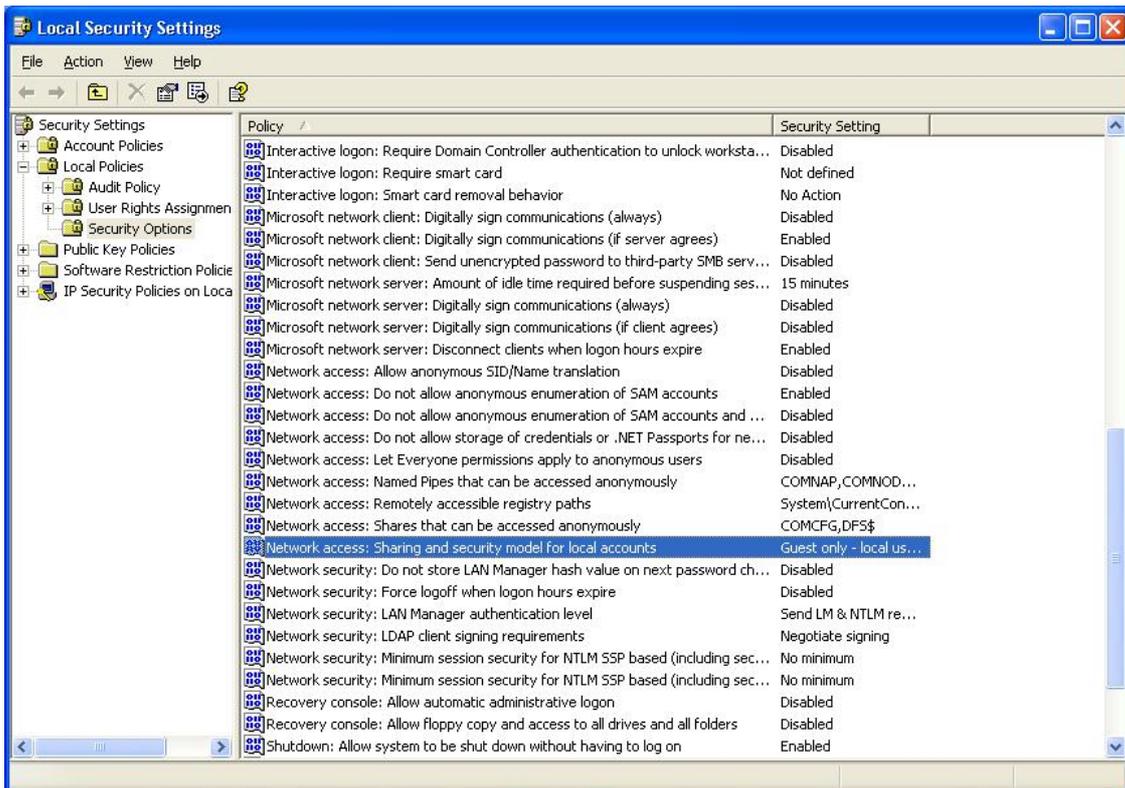


Press OK.

If all the machines in the cluster belong to a Windows Domain, the firewall configuration is now complete. Perform the following steps only if the machines belong to a Windows Workgroup.

Go to Control Panel and select Administrative tools. Choose Local Security Policy item and double-click on it. Go down the following path: Security settings > Local Policies > Security Options

On the right-side panel select *Network access: Sharing and security model for local accounts*



Double-click on the item to change the setting.



Select *Classic – local users authenticate as themselves* and press OK.
Close the Local Security Settings window.

Repeat this procedure for all the slave nodes.

Completion

If you had to stop the Mascot service to complete the firewall configuration, you can now start it again.

The installation is now complete. There will be a lot of disk activity while the Mascot service compresses the SwissProt sequence database. Searches on the database cannot be performed until the files have been compressed. You should open up the status screen in a web browser (Start menu; Programs; Mascot; Search Status) and verify the cluster status.

MASCOT search status page

Version: 2.0 - Licensed to: Matrix Science Web Site, (10 processors).
Using 5 nodes and 10 processors. [1 searches running]

[Search log](#) [monitor log](#) [error log](#) [Error message descriptions](#) [nodelist.txt](#) [Do not auto refresh this page](#)

Name = MSDB Family = C:/INETPUB/sequence/MSDB/current/MSDB *.fasta
 Filename = MSDB_20030730.fasta Pathname = C:/INETPUB/sequence/MSDB/current/MSDB_20030730.fasta
 Status = In use [Statistics](#) [Unidentified taxonomy](#)
 State Time = Tue Dec 30 11:02:19 # searches = 0
 Mem mapped = YES Request to mem map = YES Request unmap = NO Mem locked = NO
 Number of threads = 1 Current = YES

Cluster Nodes

Node	IP Address	OS	Responding	Physical Memory	Swap file	Disk space
slave03	10.0.0.3	Windows NT	☺ OK	☺ 78% free	☺ 88% free	☺ 36% free
slave01	10.0.0.1	Windows NT	☺ OK	☺ 65% free	☺ 90% free	☺ 75% free
slave04	10.0.0.4	Windows NT	☺ OK	☺ 73% free	☺ 91% free	☺ 91% free
slave11	10.0.0.11	Windows NT	☺ OK	☺ 73% free	☺ 90% free	☺ 90% free
slave33	10.0.0.33	Windows NT	☺ OK	☺ 73% free	☺ 91% free	☺ 91% free

If all is well, you will see rows of happy faces and the status line will display the following messages:

```
Creating compressed files
Running 1st test
First test just run OK
Trying to memory map files
Just enabled memory mapping
In Use
```

Once the database is “In use”, you can begin exploring and using Mascot. Clicking on the links in the cluster node table will display more detailed status information for individual nodes.

Unix

Communication

Under Unix, the master node communicates with the slave nodes using either ssh (preferred), or rsh. If communication can be established using ssh, then scp is used for file copying. If rsh is used for communication, then rcp is used for file copying.

Whether ssh or rsh is used, it is essential that communication can be established without requiring passwords or passphrases. In the case of ssh, key based authentication is the preferred mechanism. A less secure alternative for rsh is provided by file based authentication using `.rhosts` or `hosts.equiv`.

A detailed description of the many ways to configure ssh or rsh is outside the scope of this manual. For key based authentication, read the man pages for: `ssh`, `sshd`, `ssh-keygen`, `ssh-add`, `ssh-agent`. For file based authentication, read the man pages for: `rsh`, `rshd`, `rlogin`, `hosts.equiv`.

The minimum procedure to set up key based authentication for ssh on a clean Linux system, where there are no pre-existing keys, is as follows:

1. Login to the master node as the user who will own the `ms-monitor.exe` process, (generally root), and generate a version 2 RSA key pair by executing:

```
ssh-keygen -t rsa
```

2. When asked for a passphrase, press return to indicate no passphrase is required. Accept the default location for saving the key files, (`$HOME/.ssh`)
3. The contents of the public key, `$HOME/.ssh/id_rsa.pub`, must be added to a file called `$HOME/.ssh/authorized_keys` on each of the slave nodes.

4. Test communication by logging in to each slave node from the master node using ssh. The first time a connection is made, confirm that the new host should be added to the list of known hosts, `$HOME/.ssh/known_hosts`

Installation

Perform a standard installation of Mascot onto the master system according to the procedure in Chapter 2. Verify correct system operation as a single server by performing searches of MSDB and familiarise yourself with administrative tools such as `ms-review.exe` and `ms-status.exe` (Chapter 7). Any problems need to be resolved before reconfiguring for cluster operation.

Cluster Configuration Procedure

1. Kill `ms-monitor.exe`
2. Open `mascot/config/mascot.dat` in a text editor. Move down to the “Cluster” section and enter configuration information for the cluster. The parameters are fully described below in the Reference section. In the databases section, verify that the `threads` and `blocks` parameters are set to 1 for all databases. If this is not the case, make the necessary changes, then save `mascot.dat`. For a 5 node, 10 cpu cluster, typical entries might be:

```
Cluster
#
# Enable (1) or disable (0) cluster mode
Enabled 1
#
# MasterComputerName must be the hostname
MasterComputerName zx80
#
# Node defaults
DefaultNodeOS Linux
DefaultNodeHomeDir /usr/local/mascotnode
#
# Following line must be commented out WHEN this is a homogeneous
MascotNodeScript      /usr/local/mascot/bin/load_node.pl
#
# Sub-cluster definition
```

```
# Syntax is SubClusterSet X Y where X is the sub-cluster number
# and Y is the maximum number of CPUs to use within the given sub-
#
SubClusterSet 0 10
#
# Time outs, log files
IPCTimeout      5      # seconds with no response before timeout
IPCLogging      0      # no logging = 0, minimal = 1, verbose = 2
IPCLogfile      ../logs/ipc.log    # relative path
CheckNodesAliveFreq 30  # seconds between node health checks
SecsToWaitForNodeAtStartup 20  # seconds to wait for node to
#
end
```

3. Open *mascot/config/not.nodelist.txt* in a text editor. Enter configuration information for the cluster. The parameters are fully described below in the Reference section. Save as *nodelist.txt*. For a 5 node, 10 cpu cluster, typical entries might be:

```
# Cluster node definitions
#
# Each line begins with the word Node, followed by a space and
# then a comma delimited list of configuration parameters:
#   ip address:port
#   computer (host) name
#   maximum number of node CPU's to be used
#   operating system
#   local path to home directory
#   home directory as seen from master (specify for NT master only)
#
Node 10.0.0.1:5001, slave01, 2, Linux, /usr/local/mascotnode
Node 10.0.0.2:5001, slave02, 2, Linux, /usr/local/mascotnode
Node 10.0.0.3:5001, slave03, 2, Linux, /usr/local/mascotnode
Node 10.0.0.4:5001, slave04, 2, Linux, /usr/local/mascotnode
Node 10.0.0.5:5001, slave05, 2, Linux, /usr/local/mascotnode
```

4. Re-start *ms-monitor.exe*. Note that you must change directory to *mascot/bin* and have super user privileges to execute *ms-monitor.exe*.

Note: (Linux only) Under Redhat Linux 8.0, if *ms-monitor.exe* terminates immediately after launch, without any error messages, the problem may relate to a bug in *gethostbyname_r()*. In the cluster section of *mascot.dat*, try using the IP address for the master node, rather than the hostname, as the argument to *MasterComputerName*.

5. In a web browser, navigate to *ms-status.exe* and verify that the system starts up correctly.

Reference

The Cluster section in mascot.dat

```

Cluster
#
# Enable (1) or disable (0) cluster mode
Enabled 1
#
# MasterComputerName must be the hostname
MasterComputerName mascot-master
#
# Node defaults
DefaultNodeOS Windows_NT
DefaultNodeHomeDir c:/mascotnode
#
# Following line must be commented out UNLESS this is a
DefaultNodeHomeDirFromMaster \\<host_name>\c$\mascotnode
#
# Following line must be commented out WHEN this is a
# MascotNodeScript      ###ROOT###/bin/load_node.pl
#
# Sub-cluster definition
# Syntax is SubClusterSet X Y where X is the sub-cluster number
# and Y is the maximum number of processors in the sub-cluster
SubClusterSet 0 10
#
# Time outs, log files
IPCTimeout          5           # seconds with no
IPCLogging          0           # no logging = 0,
IPCLogfile          ../logs/ipc.log # relative path
CheckNodesAliveFreq 30          # seconds between node
SecsToWaitForNodeAtStartup 20     # seconds to wait for
#
end

```

Enabled

1 to enable cluster mode, 0 to enable single server mode

MasterComputerName

Enter the host name for the master computer and, optionally, the IP address separated by a comma. The IP address may need to be specified for a multi-homed master where it is necessary to define which network card is on the LAN and which is the gateway to the outside world.

DefaultNodeOS

If no OS is defined for a particular node, then this OS is assumed. Must be one of:

- Windows_NT
- Irix
- Solaris
- Linux
- AIX

Note that these names are case sensitive.

DefaultNodeHomeDir

If no specific home directory is specified for a particular node, then this default is used.

On a Unix system, this will typically be */usr/local/mascotnode*. It is best not to use */usr/local/mascot* as this is the directory mostly used for the master.

On a Windows system, this will typically be *C:/MascotNode* or *D:/MascotNode*.

To override this setting for a particular node, enter the directory on the node line

DefaultNodeHomeDirFromMaster

This is the directory on the node as seen from the master. For a Windows cluster, this must be present and specified as a UNC name.

The text `<host_name>` will be replaced by the host name as specified in the Node line.

For a Unix cluster, this parameter must be commented out.

MascotNodeScript

This script is run for each node with the following parameters:

-i ip address of node - required

-t The task to be performed - required, either

‘StopNode’ – the script will try and stop the Mascot Node daemon or service on the specified node.

or

‘StartNode’ – the script will unconditionally update *ms-mascotnode.exe*, *mascot.license*, and *mascot.dat* on the specified node, then start the Mascot Node daemon or service.

- f Full path to the node's home directory - required
- r Port number of node - required
- o The operating system running on the node – required

For a Unix cluster, the master and slave nodes must be able to communicate using either ssh (preferred), or rsh without requiring passwords or passphrases. In the case of ssh, key based authentication is the preferred mechanism. A less secure alternative for rsh is provided by file based authentication using `.rhosts` or `hosts.equiv`.

SubClusterSet X Y

Large clusters can be divided into sub-clusters. X is a unique integer value (0 based) used to identify the sub-cluster. Y is the maximum number of processors in the sub-cluster. A single cluster must have a single entry with X set to 0.

IPCTimeout

The timeout in seconds for inter-process communication

IPCLogging

- 0 for no logging of inter-process communication
- 1 for minimal logging
- 2 for verbose logging

IPCLogfile

The relative path to the inter-process communication log file

CheckNodesAliveFreq

The interval in seconds between 'health checks' on the nodes

SecsToWaitForNodeAtStartup

At startup, if a node is not available within this time, the system will continue to startup without that node. If the value is set to 0, then the system will wait indefinitely. Default is 60 (seconds).

This timeout is also used if a node fails while the system is running. The system will wait for this number of seconds before re-initialising `ms-monitor.exe`. This means that a short-lived interruption in network communication doesn't create a major service interruption.

MascotNodeRebootScript

Path to an optional CGI script to re-boot a cluster node. If this parameter is defined, there will be a link at the bottom of each Mascot Cluster Node

status page. Clicking on this link will execute the specified CGI script with the host name of the specified node as an argument.

DefaultPort

Sets the default port number to be used when this parameter is missing from *nodelist.txt*. Recommended default is 5001

UseCompleteDatabase

Not used. If specified, must be set to 1.

nodelist.txt

This file is used to define the nodes that belong to the cluster. For a very large cluster, it is advisable to define a few percent of additional nodes as 'spares'. For example, if 51 nodes with 102 processors were available, and Mascot was configured to use 2 sub-clusters, each of 50 processors, the node with the 2 spare processors could be used to replace a failed node automatically.

```
# Cluster node definitions
#
# Each line begins with the word Node, followed by a space and
# then a comma delimited list of configuration parameters:
#   ip address:port
#   computer (host) name
#   maximum number of node CPU's to be used
#   operating system
#   local path to home directory
#   home directory as seen from master (specify for NT master only)
#
Node 10.0.0.1:5001, slave01, 2, Windows_NT, c:/MascotNode, ㄿ
    ㄿ \\slave01\c$\MascotNode
Node 10.0.0.2:5001, slave02, 2, Windows_NT, c:/MascotNode, ㄿ
    ㄿ \\slave02\c$\MascotNode
Node 10.0.0.3:5001, slave03, 2, Windows_NT, c:/MascotNode, ㄿ
    ㄿ \\slave03\c$\MascotNode
Node 10.0.0.4:5001, slave04, 2, Windows_NT, c:/MascotNode, ㄿ
    ㄿ \\slave04\c$\MascotNode
Node 10.0.0.5:5001, slave05, 2, Windows_NT, c:/MascotNode, ㄿ
    ㄿ \\slave05\c$\MascotNode
```

Important: Because Mascot frequently writes status information to *nodelist.txt*, you should open the file in a text editor that puts a lock on the file (e.g. vi or wordpad). This will prevent Mascot from modifying the file while it is being edited. *nodelist.txt* can be viewed using Mascot Status.

Node

There must be one or more node entries. Items in square brackets are optional – but the commas must always be supplied.

[IP address]:Port, [Host name], Number of processors, [OS], [Home dir], [Home dir from master]

Either the IP address or the host name must be specified. The port must always be specified.

The number of processors to be used on the node can be less than the number of processors available. If the total number of processors specified in all the nodes exceeds the number of licenses available, only the licensed number will be used at any one time.

If the OS is not specified, then the DefaultNodeOS is used. Must be one of the choices shown under DefaultNodeOS.

The home directory is the local path on the node to the root of the Mascot directory structure. If this is not specified, then DefaultNodeHomeDir is used.

Home directory from master is the home directory on the node as seen from the master. This parameter is only applicable to a Windows cluster and must be omitted for a Unix cluster.

Once a cluster has been started, an additional four status values will be written periodically to *nodelist.txt*. If you edit this file while Mascot is *not* running, these values can be deleted.

subcluster ID number (0 based)

node within subcluster (0 based)

status: 0 unknown status

1 attempting to bring into use

2 no response to ping

3 failed to start service

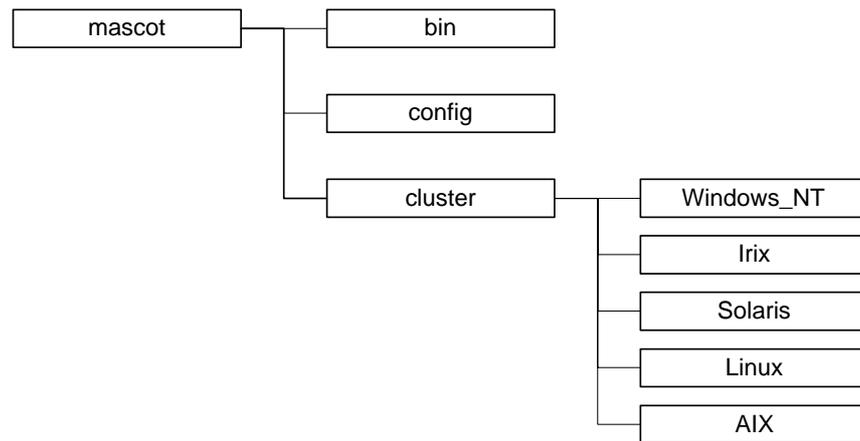
4 in use

number of CPU's actually being used

File Replication

The configuration files, such as *mascot.dat*, that are on the Mascot master are automatically replicated to the nodes. So, it is only necessary to update a file on the master. The *ms-monitor.exe* program (run as the Matrix Science Mascot Service under Windows), continually looks to

see if a file has been updated, and will distribute new versions to the nodes as required. The dates, times and lengths of the distributed files should be identical on all systems.



The same process is used for updates to executable programs, except that these updates will only be made when the *ms-monitor.exe* service first starts. Because the Mascot Cluster can be run on multiple operating systems, a distribution directory structure as shown above is used on the Mascot master server:

The Status screen will indicate if any executable files need updating.

Files required on each Mascot Node

Target File name and directory relative to node home directory
Notes

<code>./bin/ms-mascotnode.exe</code>	Updated at start-up
<code>./bin/nph-mascot.exe</code>	
<code>./config/enzymes</code>	
<code>./config/mascot.dat</code>	Updated at start-up
<code>./config/unimod.xml</code>	
<code>./config/mascot.license</code>	Updated at start-up
<code>./config/taxonomy</code>	
<code>./config/fragmentation_rules</code>	
<code>./config/quantitation.xml</code>	
<code>./taxonomy/nodes.dmp</code>	
<code>./taxonomy/usernodes.dmp</code>	Not required by most users. Note that <code>names.dmp</code> is not required on the Mascot Nodes.

Start-up of `ms-monitor.exe`

The following sequence occurs (for each node) when `ms-monitor.exe` starts for the first time on the master system. (items marked * are for Windows clusters only).

- See if the computer is available by opening a socket to the ping port (port 7)
- If there is an entry 'StopMascotNodeCmd' in the `mascot.dat` file, then run that command to stop the Mascot node daemon - or
 * See if there is a MascotNodeService installed on the computer - if there is, then stop that service
- If there is no `ms-mascotnode.exe` or if it is out of date on the Mascot node, then copy/update the file from the cluster/<OS> directory on the Mascot Master system to the specified directory on the Mascot Node.
- * If the service is not installed, then install the service, and add a registry entry for the directory to be changed to at start up
- Make a `logs` and `config` directory and copy `mascot.dat` and `mascot.license`

- * Start the MascotNodeService on the Mascot Node computer. (With a Unix based system, the *ms-mascotnode.exe* daemon will be started).
- Check that the service / daemon now communicates through TCP/IP sockets – if it fails, then a message indicating which Mascot node it is waiting for is displayed in the ms-status screen.
- Initialise the MascotNodeService / daemon by sending the appropriate commands
- See if any files are missing or out of date (see above), and if necessary, update them. This is done through the TCP/IP socket, so no directory mapping / NFS mounts are required.

Once all the Mascot nodes have been successfully initialised, then Mascot Monitor starts as normal.

Licensing

The number of processors that the search is permitted to run on is restricted by the number of mascot licenses. The Mascot master node is not included in this list, since it merely distributes the search and collates the results. The number of processors to be used for Mascot will never exceed the number specified in the licence.

Error messages and emails

In the single server version of Mascot, selected warning messages can optionally be emailed to the system administrator when something critical, such as a database update, fails on the server. The following additional messages, specific to a cluster, can also be emailed:

M00323 One or more cluster nodes has stopped responding

M00316 Dr. Watson log updated (indicating a software crash) on one of the cluster nodes.

Who Am I?

If the Mascot master is also being used as a node, when *nph-mascot.exe* is run, it needs to know whether it is running as a node task or as master task. Since the different *mascot.dat* files are identical, it determines this from a file *mascot/config/iam.dat* that is created by the Mascot node service when it starts up. Do not copy or replace this file.

Windows Manual Configuration

The following configuration steps on each slave node are performed automatically as part of the Windows installation

MascotNodeService

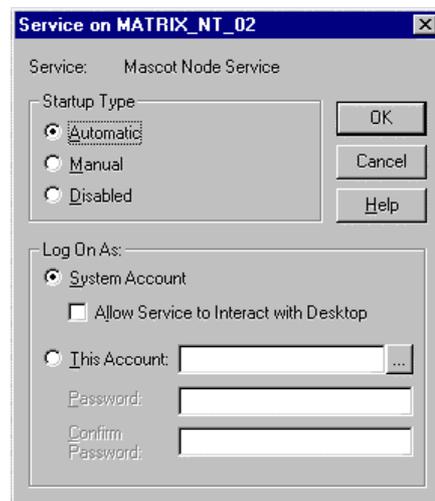
Under Windows, *ms-mascotnode.exe* is configured to run as a service. This should be taken care of automatically. If there are any problems, service creation or deletion requires the Microsoft utility *sc.exe*, which can be found in the *mascot/cluster/Windows_NT* directory.

The command to create the service is:

```
C:> sc create MascotNodeService ↵
    type= own ↵
    binpath= c:\mascotnode\bin\ms-mascotnode.exe ↵
    start= auto
```

You may need to change the path to the executable, and note that the spaces after the equals signs are significant.

To verify that the service has been created successfully, from the Control panel, open the Services control panel and choose MascotNodeService. Select Startup... and the following dialog should be displayed:



To delete the service, first stop it, then enter:

```
C:> sc delete MascotNodeService
```

Dr. Watson

To prevent (invisible) dialog boxes from being displayed if a fatal error occurs, edit the registry key

```
HKEY_LOCAL_MACHINE\Software\Microsoft\DrWatson
```

Set the value of *VisualNotification* to 0. When the Mascot node service starts on a Windows system, it sets a Dr. Watson registry entry

to ensure that Dr. Watson log files are written to the node *logs* directory.

Registry Settings

Two registry entries are used on each slave node to record the root directory of the mascot file structure and the port number used for communication. For example:

```
[HKEY_LOCAL_MACHINE\SOFTWARE\MatrixScience\Mascot\1.00]
"MascotNodeFolder"="C:/mascotnode/bin"
"MascotNodePort"="5001"
```

12

Security

Overview

The security model allows a Mascot administrator to:

- Prevent un-authorised changes of Mascot server configuration files using, for example, the database maintenance utility
- Restrict access to results files and sequence databases based on group and user definitions
- Provide standard 'session' support (with time-outs) so users do not need to continually re-enter passwords
- Restrict access to Mascot server based utilities that allow deletion of searches and other job control functions
- Provide read-only access to configuration files for third party applications without requiring login
- Optionally allow submission of searches etc. for 3rd party applications without a login
- Switch OS platform painlessly if Mascot or Mascot Integra authentication is used
- Easily set-up Mascot Daemon to run searches as the 'customer' in a service or core lab environment

Enabling security

When Mascot is first installed, the security system is disabled. To enable security, open a command prompt or shell on the Mascot server and change to the `mascot/bin` directory. Enter the command:

```
perl enable_security.pl
```

The Mascot service (`ms-monitor.exe`) must then be stopped and re-started.

Disabling security

To disable security, open a command prompt or shell on the Mascot server and change to the mascot/bin directory. Enter the command:

```
perl disable_security.pl
```

The Mascot service (ms-monitor.exe) must then be stopped and restarted.

Authentication

There are two different ways in which users can be authenticated:

1. Mascot authentication. The passwords are stored and maintained by the Mascot security libraries and/or by Mascot Integra.
2. Web server authentication. Available with any web server that supports authentication. Refer to your web server documentation for details on how to set up authentication

The type of authentication is set up at the user level, and not as a global setting. Even if the server has web authentication switched on, it may be useful to set some users to be authenticated using the Mascot authentication. A typical case for this might be for a service lab manager running Windows and IIS with integrated authentication. This user would not typically want to create a separate Windows login account for the administrator, but would choose to login explicitly as administrator to update configuration files etc. For an Apache server, with authentication switched on, most users would want to be set to use the authenticated login

Users

New users are added using the Mascot security administration utility. There are 6 special “system” user accounts:

guest

The guest user is not enabled by default. If this account is enabled, then any user is automatically logged in as guest, and needs to explicitly login as another user to gain further access rights. The guest account cannot be deleted, but the account can be disabled. The userid is 1.

admin

This account should be used to perform administration on the Mascot server. It is recommended that you always log in as administrator to perform security and other administration rather than assign administrator rights to another user. The administrator account cannot be

deleted or disabled and the admin user cannot be removed from the administrators group. By default, the administrator can access all the administrator screens, but cannot submit searches. The userid is 2. The initial password for admin is admin, but this must be changed on first login.

Command line

This pseudo user is always used when running programs from the command line, and can perform any task without restriction. This 'user' doesn't appear in the security administration utility and hence the account cannot be deleted or disabled. The userid is 3.

daemon

This user should be used to run searches in Mascot Daemon. See the Mascot Daemon help for details. The user account is disabled by default, so it will need to be enabled and before use. The userid is 4.

public_searches

This is a pseudo user that is used for the example searches. This 'user' doesn't appear in the security administration utility and hence the account cannot be deleted or disabled. It isn't possible to login as this user. The userid is 5.

(system)

The Mascot Integra system account is used to query data on the Mascot server. Do not change the name of this account or the type of the account. There is no password associated with this account since it can only be called from the secure Mascot Integra server. The userid is 6.

Types of user

Six 'types' of user are available, and the appropriate type should be selected using a the drop down list in the administration screen:

Standard Mascot User

The user name and password are stored by Mascot

Mascot Integra User

The password, password expiry and timeouts for these users are set in the Mascot Integra administration screens.

The standard Mascot login screen will be displayed, but authentication is performed using Mascot Integra

The Mascot Integra server details must be specified in the options section of the security administration utility.

IP address

This 'user' should only be used for third party legacy applications, that do not support Mascot security. Instead of a user name, enter the static IP address of the computer that will access the Mascot server. Do not enter a password.

Computer name

Same as the IP address, but the computer name is used instead. A computer name is more practical where dynamic IP addresses are being used.

Agent string

Should only be used as a last resort for third party applications that haven't implemented Mascot security and where the computer name / IP address is not reliable. A case sensitive substring comparison will be made with the HTTP_USER_AGENT environment variable.

Use built in web server authentication

See description of 'authentication' above.

Mascot will never prompt these users for username and password, and hence passwords and password expiry will be ignored.

Mascot security session time-outs do not apply.

In Microsoft Internet Information Services, (IIS), if anonymous access and integrated authentication are both enabled, then users will generally be 'logged in' as anonymous until they try to access a file where permission is denied. This almost certainly means that anonymous login must be disabled to use this option.

IIS user names generally include the Domain name: e.g. matrix_science/charles. The comparison will be with everything after the last forward or back slash. So, in this case, you would enter 'charles' as the user name.

Groups

Access rights can be assigned to groups, not users. Therefore, a user has no effective rights unless they belong to one or more groups. If a user belongs to more than one group, then their rights are the combination of the rights in both groups.

There are 5 special built in groups:

Guests

By default, the guest user is the only member of this group and the guest group can only submit PMF searches against any database. This can easily be changed using the security administration utility.

Administrators

The admin user always belongs to this group. Members of the group can perform any administration task, but cannot submit searches.

PowerUsers

Members of this group can submit all types of searches and perform some administration. They cannot access the security administration utility.

Daemons

The daemon user belongs to this group by default.

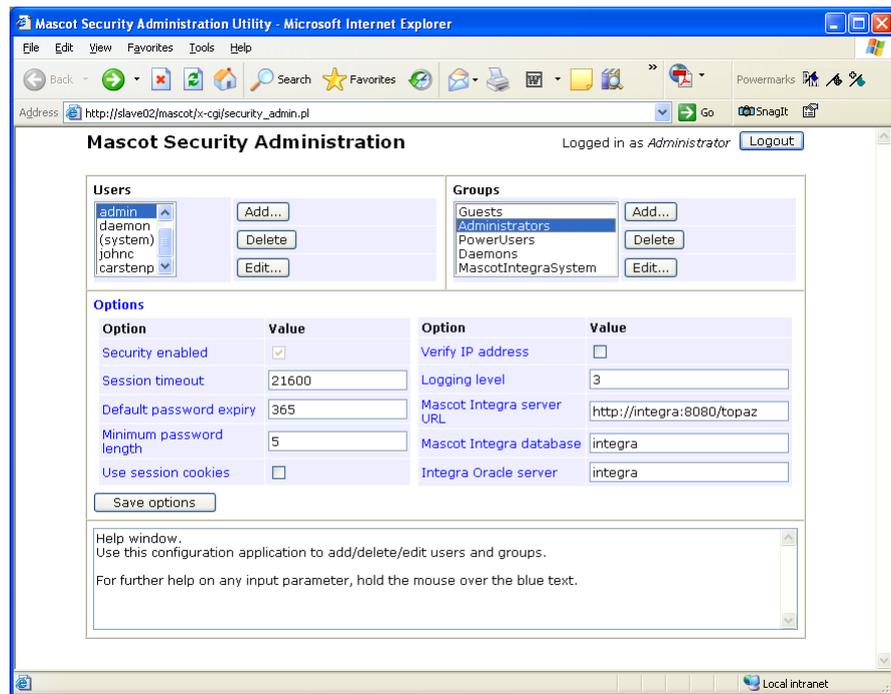
MascotIntegraSystem

The (system) user is the only member of this group.

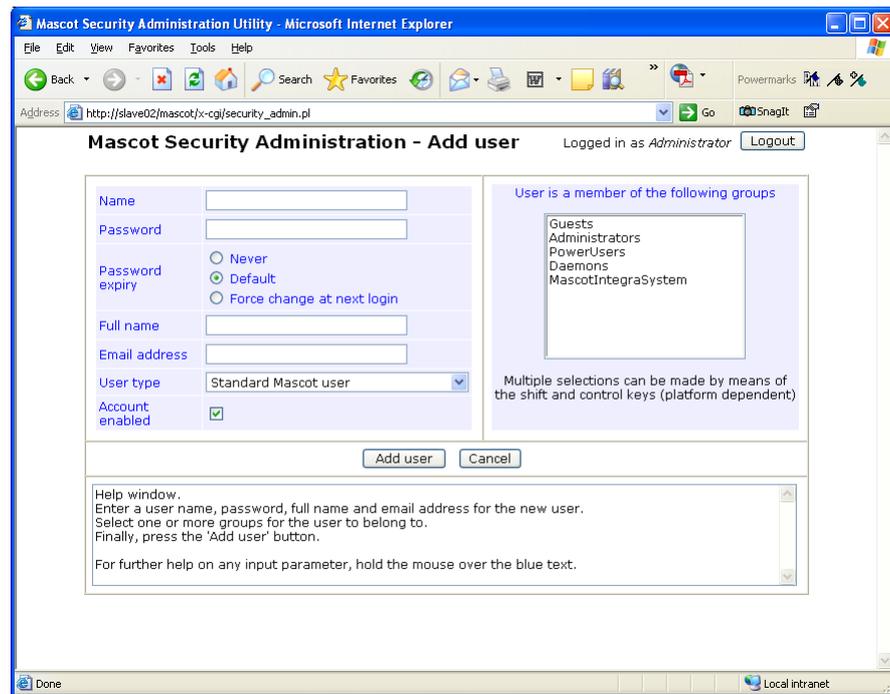
Using the security administration utility

When the security administration utility is started for the first time, you will need to login as admin/admin. You are then forced to change the password.

The main page lists the current users and groups and has buttons for deleting adding and editing users and groups. The global security options can also be modified from this page. On all the pages, there is a help window that gives details about specific options – just position the mouse over the relevant hyperlink to see the help.

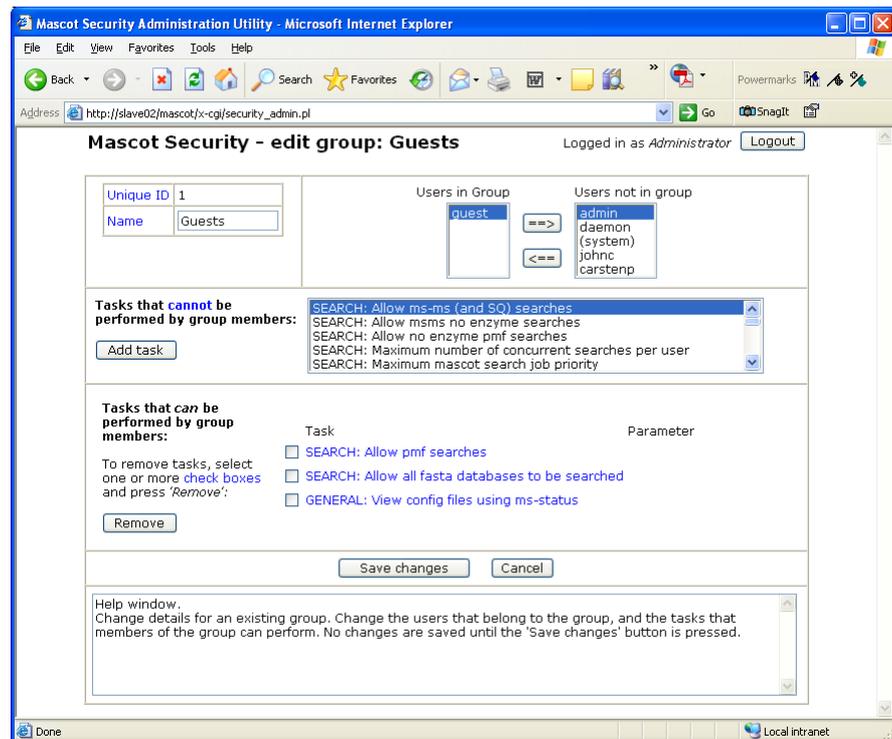


To add a new user, click on the Add... button:



The new user must be given a name, password, full name and email address. There is a description of the different user types of user earlier in this chapter. You should also select one or more groups that the user should belong to before pressing the 'Add user' button.

New groups may be added or edited:



Each group has a unique ID that cannot be changed.

Users can be added to or removed from groups either on this screen or from the edit/add user screens.

Mascot security is fine grained. There is a list of about 20 tasks that members of a group can (or cannot perform). The tasks that are not permitted are in the top list. To allow group members to perform one of these tasks, click on the task in the list, and then 'Add task'. This task will then appear in the lower list. Similarly, to remove a task, click on the check box in the lower list, and click on the 'Remove' button. To get further information about any task, hold the mouse over the task in the lower window and further details will appear in the help box.

No changes to a group are saved until the 'Save changes' button is pressed.

Session files

Session files are created in the mascot/sessions directory. Sessions that have expired will be deleted automatically by ms-monitor.

Log file

The log file 'security.log', in the mascot/logs directory contains information about all security changes. The file is not available from any web based application for security reasons. The level of logging can be controlled from the security administration utility.

Configuration Files

Security information is saved in three configuration files in the mascot/config directory:

```
security_options.xml
security_tasks.xml
group.xml
user.xml
```

The schema for these files is mascot_security_1_0.xsd.

Use the security administration utility or Mascot Parser rather than editing these files manually.

Automating addition of new users

Mascot Parser users have access to all of the documentation for the lower level functions to administer Mascot security programmatically. The security administration utility uses some of these functions.

To simply to add a large number of users, then the add_user.pl script in the mascot/bin directory can be used:

```
Usage: add_user.pl -u username
                    -p password
                    -x password_expiry
                    -f fullname
                    -e email_address
                    -g group to which user should belong
```

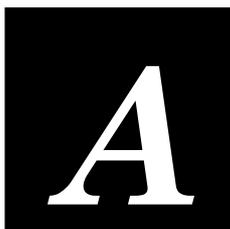
The password expiry should be 0 for never expires or 1 to force the user to change the password when they first log in.

Resetting the administrator password

If the admin user password is lost, the easiest way to reset it is to re-run 'enable_security.pl' from the command line as described above. This will not affect any existing groups or users, but will just reset the password.

User ID

The user ID for each search is saved in the results file. If security is disabled, then the search ID will be set to zero. Special user IDs are listed above. Other users will have an automatically assigned IDs starting at 1000.



Basic Regular Expressions

Sequence database parsing in Mascot is defined using rules which conform to Basic Regular Expression (BRE) notation as defined in standard ISO/IEC 9945-2: 1993. BRE notation is widely used in Unix, e.g. in the `grep` command, but it may be less familiar to those from a DOS or Windows background. Man pages containing a rigorous definition of BRE notation can be found on most Unix systems.

The following description is much simplified, and is intended to provide just enough information to understand the existing rules in `mascot.dat`, and to enable someone without prior knowledge of regular expressions to write simple rules for new databases. Only the most basic aspects of BRE notation are touched on.

In `mascot.dat`, the `PARSE` section contains a number of rules. For each rule, the pattern in double quotes is a BRE which is used to identify a string so that it can be parsed from the surrounding text. For example:

```
#Report text from NCBI excluding sequence (used for AA entries)
RULE_10 "\(LOCUS .*\)ORIGIN "
```

The part of the BRE between the backslashed parentheses `\(` and `\)` is the string which we are trying to locate and extract. This rule looks for the word `LOCUS` followed by a space. It will extract all the text, including the word `LOCUS`, up to but excluding the word `ORIGIN` followed by a space.

BRE Rules

The rules for performing this match are as follows:

The BRE always looks for the longest, leftmost matching string.

Matching is case sensitive.

Newline characters (LF in Unix or CR+LF in Windows) are treated like any other character

The sub-expression to be extracted from the surrounding text is defined using backslashed parentheses `\(\)`. The parentheses are ignored for matching purposes.

Some characters are “Special”:

- `.` `[` `\` The period, left-bracket and backslash are special except when used in a bracket expression.
- `*` The asterisk is special except when used in a bracket expression, as the first character of an entire BRE (after an initial `^`, if any), as the first character of a subexpression (after an initial `^`, if any).
- `^` The circumflex is special when used as an anchor, or as the first character of a bracket expression.
- `$` The dollar sign is special when used as an anchor.

Matching Single Characters

Any character that is not a special character is an ordinary character. An ordinary character, or a special character preceded by a backslash, matches to itself.

A period, used outside a bracket expression, matches to any single character, including a newline character.

A bracket expression (a list of characters enclosed in square brackets, `[]`) matches any single character from the enclosed list. The following rules and definitions apply to bracket expressions:

A bracket expression is either a matching list expression or a non-matching list expression. The right-bracket `]` loses its special meaning and represents itself in a bracket expression if it occurs first in the list (after an initial circumflex `^`, if any). Otherwise, it terminates the bracket expression. The special characters: `.` `*` `[` `\` (period, asterisk, left-bracket and backslash, respectively) lose their special meaning within a bracket expression.

A matching list expression matches any one of the characters in the list. The first character in the list must not be the circumflex. For example, `[abc]` matches any one of the characters `a`, `b` or `c`.

A non-matching list expression begins with a circumflex `^` and specifies a list that matches any character except for the characters in the list after the leading circumflex. For example, `[^abc]` matches any one character except the characters `a`, `b` or `c`. The circumflex will have this special meaning only when it occurs first in the list, immediately following the left-bracket.

A range expression represents the inclusive set of characters between two characters in the ASCII character set. The starting and ending characters are separated by a hyphen. For example, `[A-Z]` will match to any single upper case letter, while `[0-9_A-Za-z]` matches any single alphanumeric character.

Matching Multiple Characters

When a BRE matching a single character or a subexpression is followed by the special character asterisk `*`, together with that asterisk it matches what zero or more consecutive occurrences of the character. For example, `[ab]*` and `[ab][ab]` are equivalent when matching the string `ab`. The expression `ab*c` will match to `ac` or `abc` or `abbbbbbc`.

When a BRE matching a single character or a subexpression is followed by an interval expression of the format `\{m\}`, `\{m,\}` or `\{m,n\}`, together with that interval expression it matches what repeated consecutive occurrences of the BRE would match. The values of `m` and `n` will be decimal integers in the range $0 \leq m \leq n \leq 255$, where `m` specifies the exact or minimum number of occurrences and `n` specifies the maximum number of occurrences. The expression `\{m\}` matches exactly `m` occurrences of the preceding BRE, `\{m,\}` matches at least `m` occurrences and `\{m,n\}` matches any number of occurrences between `m` and `n`, inclusive.

For example, in the string `abababcccccccd` the BRE `c\{3\}` is matched by characters seven to nine, the BRE `\(ab\)\{4,\}` is not matched at all and the BRE `c\{1,3\}d` is matched by characters ten to thirteen.

The behaviour of multiple adjacent duplication symbols produces undefined results.

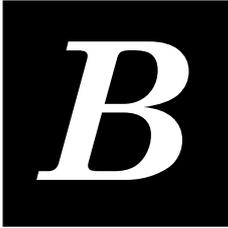
Expression Anchoring

A BRE can be limited to matching strings that begin or end a line; this is called anchoring. The circumflex and dollar sign special characters will be considered BRE anchors in the following contexts:

A circumflex `^` is an anchor when used as the first character of an entire BRE. The circumflex will anchor the expression to the beginning of a string; only sequences starting at the first character of a string will be matched by the BRE. For example, the BRE `^ab` matches `ab` in the string `abcdef`, but fails to match in the string `cdefab`.

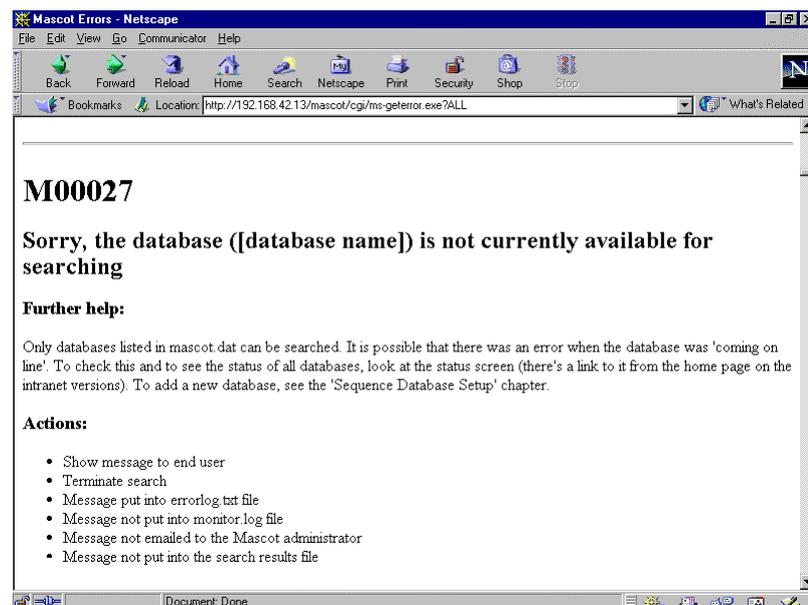
A dollar sign `$` is an anchor when used as the last character of an entire BRE. The dollar sign will anchor the expression to the end of the string being matched (not including a final newline character, if present).

A BRE anchored by both `^` and `$` matches only an entire string. For example, the BRE `^abcdef$` matches strings consisting only of `abcdef`.

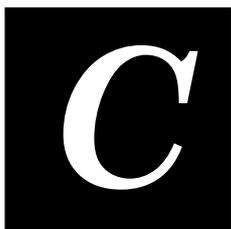


Error Messages

A complete listing of Mascot error codes, messages, and explanations can be found at the URL `mascot/cgi/ms-geterror.exe?ALL`.



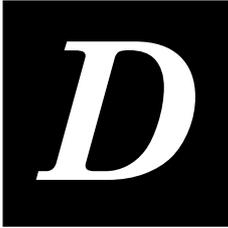
The same text can also be found in the file `errors.html` in the root directory of the Mascot CD-ROM.



System Limits

Number of different modifications in mod_file	unlimited
Number of enzymatic peptides per sequence	user definable (MaxNumPeptides)
Length of a sequence (number of residues)	user definable (MaxSequenceLen)
Number of seq(), comp(), and ions() type qualifiers per query	20
Maximum number of tags and etags in a search	100
Number of peptide masses (MS/MS search)	unlimited
Number of peptide masses (PMF search)	1000
Number of enzymes in the enzymes file	100
Number of protein hits saved in the results file summary section (PMF)	50
Number of peaks per MS/MS spectrum	10,000
Number of lines with name= in MIME format file	1,000,000
Maximum mass of any peptide (Daltons)	16,000
Minimum mass of any peptide (Daltons)	100
Length of any peptide in residues	254
Length of name (TITLE=) for any query when 'escaped'	30,000
Length of database name	19
Length of enzyme name	50
Length of modification name	50
Simultaneous variable modifications	9
Number of active sequence databases	64
Number of threads per search	1024
Number of concurrent jobs per database	100
Number of parse rules	256

Length of parse rule	128
Maximum number of processors per server	64
Number of missed cleavage sites in a peptide	9
Maximum number of sub-clusters in a cluster	50
Maximum number of machines in a sub-cluster	1024
Maximum number of processors in a sub-cluster	65536
Maximum number of nodes in nodelist.txt	4096



Web Server Configuration

Mascot Directory Structure

There are two directory structures to consider. One consists of the “real” paths to files on disk, the other consists of the “virtual” directories which define the web server URL’s. The virtual directories are mapped to real directories.

Any directory which contains CGI executable programs (e.g. *nph-mascot.exe*) or scripts (e.g. *master_results.pl*) must have execution access enabled.

Under normal circumstances, if a directory is mapped to a URL, all of its subdirectories are also accessible as subdirectories of the URL. Figure D.1 shows the recommended directory structure for Mascot. The root of this structure can be any convenient path.

Some of the directory paths can be changed by modifying the configuration file, *mascot.dat*. For example, it may be desirable to have the FASTA database files on a separate disk from the rest of the files. Care should be taken with any changes which affect a URL mapped directory or file, because this may require one or more HTML files to be edited to modify links.

In most cases, the contents of the directories can be deduced from their names:

bin contains (non-CGI) executables.

cgi contains CGI executables

cluster contains one or more sub-directories for platform specific executables, for distribution to the nodes in a cluster

config contains configuration files

data contains Mascot results files. By default, a new sub-directory is created for each day’s results files. The name of each sub-directory is that day’s date in ISO format, *yyyymmdd*.

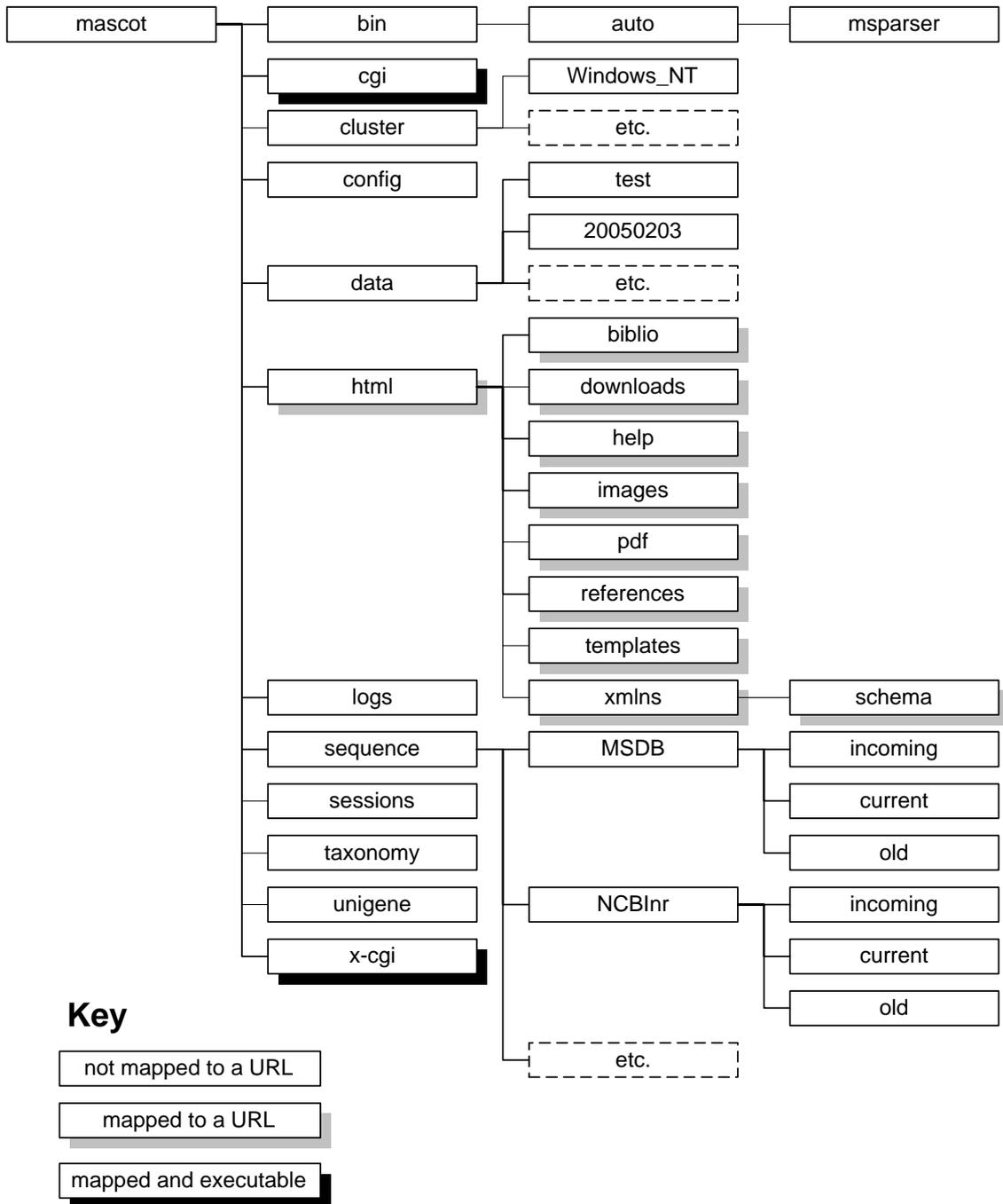


Figure D.1 Mascot Directory Structure

html is the root directory for documents

logs contains search and error logs, etc.

sequence contains a sub-directory for each FASTA database. As illustrated, for each database there are 3 sub-directories to organise the FASTA files into new downloads (*incoming*), active databases (*current*) and the most recently replaced files (*old*).

sessions contains security session files

taxonomy contains taxonomy resources

unigene contains sub-directories for species specific UniGene indexes

x-cgi is a directory for administrative CGI executables, to which access may need to be restricted. This can be achieved using either Mascot security or web server security.

Microsoft Internet Information Services

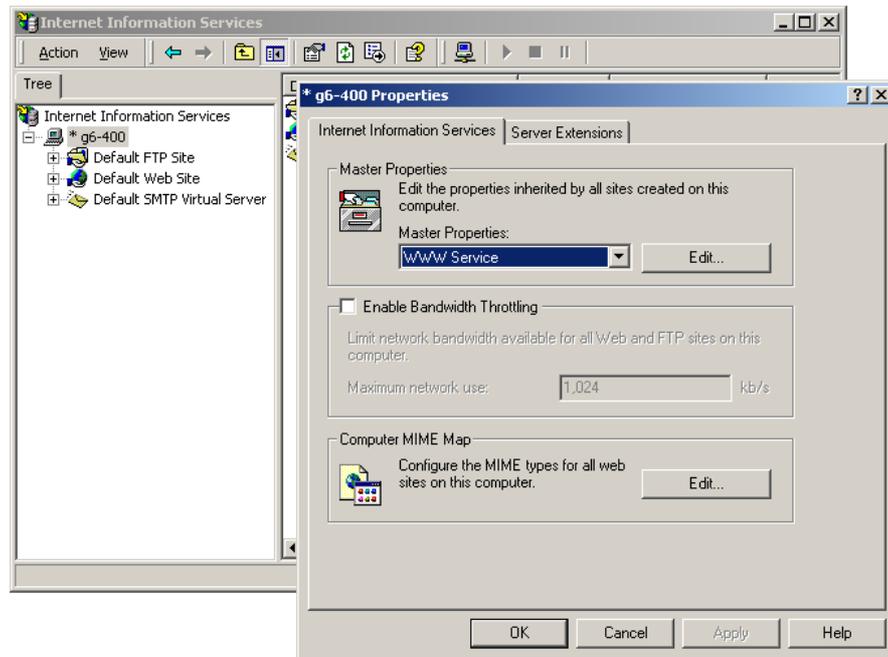
The Mascot installation program automatically configures Microsoft IIS 4.0 or later. The CGI timeout is set to 1 day, and any searches running longer than this will be terminated. If you need to increase this timeout, you may be able to use the IIS Management Console. If not, it can be modified at a command prompt or by using the MetaEdit utility, as described below.

CGI Timeout

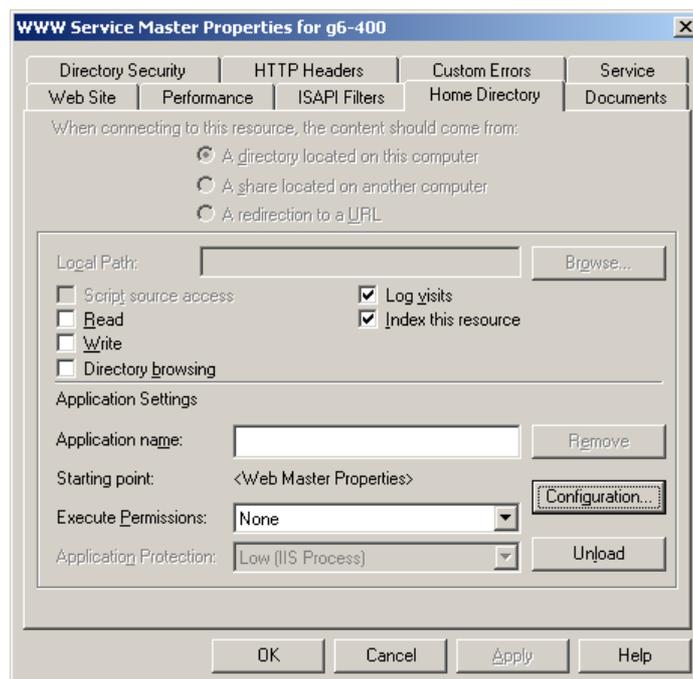
IIS Management Console

Using the IIS Management Console, it is only possible to set the CGI timeout at a 'global' level. (The Mascot installation sets it for just the Mascot directories).

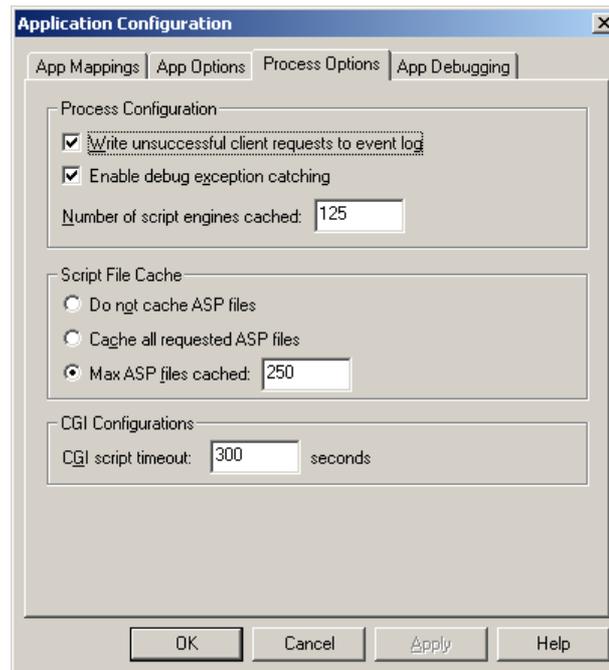
From the Control Panel, Administrative Tools, launch the IIS Management Console, (may be listed as Internet Information Services). Right click on the computer name. (This is **very important** – do not right click on "Default Web Site", which is more logical). Select 'Properties' from the menu.



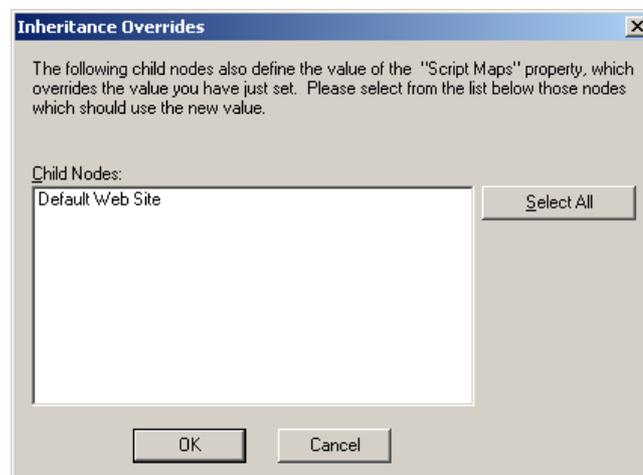
Click on the 'Edit...' button, and then the 'Home Directory' tab:



and then on the 'Configuration...' button:



The 'Process Options' tab is only available if you go through this route. The CGI script timeout is at the bottom of the dialog box and is in seconds. The Mascot installation default is set to $24 * 60 * 60 = 86400$. If you enter (say) 100000 and click apply, the following may be displayed:



Select the “Default Web Site” and press OK.

Command Prompt

At the command prompt, go to `c:/Inetpub/AdminScripts` directory

To get the value of the current Mascot cgi timeout:

```
cscript adsutil.vbs get /w3svc/1/root/mascot/cgi/cgitimeout
```

The default value as set by Mascot will be 86400 seconds

To set a new timeout value:

```
cscript adsutil.vbs set /w3svc/1/root/mascot/cgi/cgitimeout x
```

where x is the value in seconds that you want to set

Restart IIS

MetaEdit

(Not for Server 2003 and Vista)

Download and install Microsoft MetaEdit 2.2

XP: <http://support.microsoft.com/kb/232068>

NT4 or 2000: <http://support.microsoft.com/kb/301386>

By default, it will be installed under Administrative Tools. Launch MetaEdit, and expand the tree

LM > W3SVC > 1 > ROOT > mascot > cgi

You should then find an entry for CGITimeout. Change this to the required value in seconds. 1 day = 24 * 60 * 60 = 86400

Apache

Apache is a very rugged and popular server for Unix platforms. It is a less obvious choice for Windows, since the Mascot installation program will configure Microsoft IIS automatically.

Note: If the URL `/mascot` is mapped to disk path `mascot/html`, then URL `/mascot/images` will correspond to disk path `mascot/html/images`. So, it is important that the entries for the `cgi` and `x-cgi` directories come before that for the `html` directory. Otherwise, the server will report that it cannot find the `cgi` and `x-cgi` paths, because it has assumed from the URL that they are sub-directories of `mascot/html`.

Unix configuration

From Apache version 1.3.6 onwards, all configuration directives go into the file *httpd.conf*. In earlier versions, the files *access.conf* and *srm.conf* were also used.

The following lines illustrate how to configure mappings and permissions for the Mascot directories:

```
<Directory /usr/local/mascot/cgi>
  AllowOverride None
  Options None
  Order allow,deny
  Allow from all
</Directory>
ScriptAlias /mascot/cgi /usr/local/mascot/cgi

<Directory /usr/local/mascot/x-cgi>
  AllowOverride None
  Options None
  Order allow,deny
  Allow from all
</Directory>
ScriptAlias /mascot/x-cgi /usr/local/mascot/x-cgi

<Directory /usr/local/mascot/html>
  AllowOverride None
  Options None
  Order allow,deny
  Allow from all
</Directory>
Alias /mascot /usr/local/mascot/html
```

Windows Installation

If you choose to use Apache under Windows, a good starting point for support information is:

<http://httpd.apache.org/docs-2.0/platform/windows.html>

Important: If IIS is installed, stop the IIS service before installing Apache, otherwise you will need to un-install and re-install Apache.

Install Apache. Then, from the Windows Start menu, choose Programs; Apache HttpServer 2.0.x; Configure Apache Server.

The Mascot installation program creates a file containing the required configuration directives and saves it to the Mascot config directory. You simply need to paste this text into *httpd.conf* and save it.

From the Windows Start menu, choose Programs; Apache HttpServer 2.0.x; Control Apache Server; Restart.

Important: In *mascot.dat*, ensure that `ForkForUnixApache` is set to 1

Using shebang under Windows

The configuration file created by the installer includes this directive:

```
ScriptInterpreterSource Registry
```

This enables Windows-style registry file associations. Assuming Perl has been installed correctly, the extension *pl* will be associated with *perl.exe*.

Without this directive, Apache uses the shebang line at the top of each Perl script to associate the script with the Perl interpreter. The default shebang line is:

```
#!/usr/local/bin/perl
```

If you want to use this on a Windows system, you will need to change the shebang lines of all scripts to something similar to:

```
#!c:/perl/bin/perl.exe
```

User authentication.

Apache provides several ways to restrict access to directories or files. One method is to limit access to clients from a range of IP addresses or a particular domain. Another method is to require a username and password, which may be a convenient way for a system administrator to limit access to the *x-cgi* directory.

Setting up user authentication takes two steps: firstly, creating a file containing the usernames and passwords. Secondly, telling the server what resources are to be protected and which users are allowed (after entering a valid password) to access them.

Creating a User Database

A list of users and passwords needs to be created in a file. For security reasons, this file should not be under the document root. This example assumes the file is called */usr/local/mascot/config/.passwd*.

The file will consist of a list of usernames and a password for each. The format is similar to the standard Unix password file, with the username and password being separated by a colon. However you cannot just type in the usernames and passwords because the passwords are stored in an encrypted format.

The program *htpasswd* is used to add create a user file and to add or modify users. This can be found in the *bin* directory of the Apache

distribution. To create a new user file and add the username “mickey”, the command would be:

```
./htpasswd -c /usr/local/mascot/config/.passwd mickey
```

The `-c` argument tells `htpasswd` to create new users file. You will be prompted to enter a password for mickey, and confirm it by entering it again. Other users can be added to the existing file in the same way, except that the `-c` argument is not needed. The same command can also be used to modify the password of an existing user.

Specifying the password protected resources

Having created a password file, the next step is to modify the configuration mapping for the `x-cgi` directory. Instead of the mapping shown earlier, you would use a directive like this:

```
<Directory /usr/local/mascot/x-cgi>
  AllowOverride None
  Options None
  AuthType Basic
  AuthName Restricted
  AuthUserFile /usr/local/mascot/config/.passwd
  require valid-user
</Directory>
ScriptAlias /mascot/x-cgi /usr/local/mascot/x-cgi/
```

You will need to stop and restart Apache, or send a `kill -HUP` to the parent process, to activate the new configuration. For further information on restricting access to the server, see the “Authentication and Access Restrictions” section of the Apache FAQ documentation.

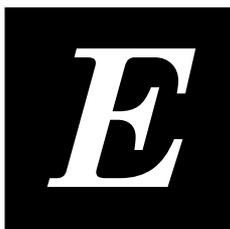
Troubleshooting

No progress reports during search

Unbuffered output mode is broken in versions of Apache 2.0 prior to 2.0.44. This prevents continuous progress reports being displayed during a search. The solution is to upgrade to a later version of Apache

<http://rhn.redhat.com/errata/RHSA-2003-139.html>

http://bugzilla.redhat.com/bugzilla/show_bug.cgi?id=82587



Web Browsers

The same version of a browser running on different platforms can behave very differently. It is virtually impossible to test all browser and platform combinations, so we restrict our testing to version 4 and later of the mainstream products:

Konqueror	http://www.konqueror.org/
Microsoft Internet Explorer	http://www.microsoft.com/windows/ie/downloads/default.asp
Mozilla	http://www.mozilla.org/
Netscape Navigator	http://home.netscape.com/download/index.html
Opera	http://www.opera.com/
Safari	http://www.apple.com/macosx/features/safari/

Compatibility with Mascot

The easiest way to check the compatibility of a specific browser is to use the “browser check” link from the Mascot home page. The functional requirements are:

- JavaScript is used widely:
 - To verify the entries in the search form fields
 - To provide an animated overview of search results
 - To save custom settings
 - To display match cross reference information in the browser status bar when the mouse cursor is moved over a query hyperlink in the peptide summary report
- Tables are used in many places
- File upload is required to submit a data file to Mascot

- Cookies are used to save user name and email address for the search form
- Style sheets are used to display a pop-up window containing the top ten peptide matches when the mouse cursor is moved over a query hyperlink in a peptide summary report

Known problems with specific browsers

Internet Explorer 5.5 (Windows)

The loading of a very large search result page is much slower in MSIE 5.5 than in either 5.0 or 6.0. If you routinely do large searches, you probably want to upgrade or downgrade.

Konqueror 3.x (various Unix)

In the Mascot search form, the default for the MASS radio button cannot be set to Average.

Opera 5.x (Windows)

Setting Mascot search form defaults to display the complete list of modifications has no effect.

File upload button has to be clicked twice to fire