GeneMine™

Discovery Engine USER GUIDE + TECHNICAL REFERENCE MANUAL

Molecular Applications Group
607 Hansen Way, Bldg. 1
Palo Alto, CA 94304

POB 51110, Palo Alto, CA 94303.1110
TEL: 650-846-3570
USA only: 1-800-229-7382
FAX: 650-846-3595
E-MAIL: support@mag.com
The GeneMine™ program, manual and associated data files are copyrighted by Molecular Applications Group. Under the copyright laws, this manual cannot be reproduced in any form without prior written permission of MAG. No patent liability is assumed, however, with respect to the use of the information contained herein.

Copyright © 1992-1997 by Molecular Applications Group, Stanford University and Yeda. All rights reserved.

This manual has been written with care, validated against the Look program and checked for accuracy. Nevertheless, Molecular Applications Group assumes no liability for damages incurred directly or indirectly from errors, omissions, or discrepancies between the program, the data files and this manual.

GeneMine™


W3C software is copyright © 1995 by: Massachusetts Institute of Technology (MIT), CERN

This W3C software is being provided by the copyright holders under the following license. By obtaining, using and/or copying this software, you agree that you have read, understood, and will comply with the following terms and conditions: Permission to use, copy, modify, and distribute this software and its documentation for any purpose and without fee or royalty is hereby granted, provided that the full text of this NOTICE appears on ALL copies of the software and documentation or portions thereof, including modifications, that you make. THIS SOFTWARE IS PROVIDED "AS IS," AND COPYRIGHT HOLDERS MAKE NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR IMPLIED, BY WAY OF EXAMPLE, BUT NOT LIMITATION, COPYRIGHT HOLDERS MAKE NO REPRESENTATIONS OR WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE OR THAT THE USE OF THE SOFTWARE OR DOCUMENTATION WILL NOT INFRINGE ANY THIRD PARTY PATENTS, COPYRIGHTS, TRADEMARKS OR OTHER RIGHTS. COPYRIGHT HOLDERS WILL BEAR NO LIABILITY FOR ANY USE OF THIS SOFTWARE OR DOCUMENTATION.

GeneMine is a registered trademark of Molecular Applications Group
Imdad and Bioscape are registered trademarks of Molecular Applications Group.
SEGMOD, ENCAD, CARA, MacImdad and PCImdad are trademarks of Molecular Applications Group.
FLEXlm is a trademark of Globetrotter Software, Incorporated.
eXodus is a trademark of White Pine Software Incorporated.
Macintosh is a trademark of Apple Computer Incorporated.

Netscape is a trademark of Netscape Communications Corporation.

Postscript is a trademark of Adobe Incorporated.

ReadSeq is copyright © 1990 by D. G. Gilbert.

The name and trademarks of copyright holders may NOT be used in advertising or publicity pertaining to the software without specific, written prior permission. Title to copyright in this software and any associated documentation will at all times remain with copyright holders.
## Contents

1. **User Guide: Discovery Engine**  
   - User Guide chapter contents  
   - Introduction  
   - About the User Guide  
   - Conventions  
   - Database licensing  
   - Discovery Engine overview  
   - Beginning a Discovery Engine project  
   - Discovery Engine’s automated annotations  
     - Annotations are sequence-associated  
     - Where annotations come from  
     - Calculated homology annotations  
     - Annotation scoring & uncertainty  
     - Homolog clusters  
     - Standard annotations  
   - Discovery Engine annotations  
     1. Domain  
     2. Active site  
     3. Secondary structure  
     4. Sequence homologs *  
     5. Sequence polymorphic regions *  
     6. Specificity Pocket  
     7. Epitope  
     8. Functional motifs  
     9. Structural homologs *  
     10. Post-translational modifications  
     11. Numbering  
     12. Closely related homologs*
## Contents

13. EST Expression* - - - - - - - - - - - - - - - - - - - - - - - - - - - -13
14. Metabolic pathway/enzyme activity - - - - - - - - - - - - - - - - - - -14
15. Genetic map/linkage - - - - - - - - - - - - - - - - - - - - - - - - -14
16. Disease Association - - - - - - - - - - - - - - - - - - - - - - - - -14
17. Homology Fingerprint* - - - - - - - - - - - - - - - - - - - - - - - - -15
18. Restriction Sites - - - - - - - - - - - - - - - - - - - - - - - - - - - -16
19. Open reading frame - - - - - - - - - - - - - - - - - - - - - - - - - -16
20. PCR primer prediction - - - - - - - - - - - - - - - - - - - - - - - - -17
21. Disulfide bridges - - - - - - - - - - - - - - - - - - - - - - - - - - - -17
22. Annotations & literature - - - - - - - - - - - - - - - - - - - - - - - -17
    Annotation format & display - - - - - - - - - - - - - - - - - - - - - - -18
    User-defined search terms- - - - - - - - - - - - - - - - - - - - - - - -18
    Hotlist annotations - - - - - - - - - - - - - - - - - - - - - - - - - - -18
    Disease association annotations - - - - - - - - - - - - - - - - - - - -19
    Expression annotations - - - - - - - - - - - - - - - - - - - - - - - - -19
    User-created annotations - - - - - - - - - - - - - - - - - - - - - - - -20

Discovery Engine Setup - - - - - - - - - - - - - - - - - - - - - - - - - -20
    Workup levels - - - - - - - - - - - - - - - - - - - - - - - - - - - - -20
      No workup - - - - - - - - - - - - - - - - - - - - - - - - - - - - -22
      Light workup- - - - - - - - - - - - - - - - - - - - - - - - - - - - -22
        Protein sequences - - - - - - - - - - - - - - - - - - - - - - - - -22
        DNA sequences - - - - - - - - - - - - - - - - - - - - - - - - - -22
        Protein sequences - - - - - - - - - - - - - - - - - - - - - - - - -23
        DNA sequences - - - - - - - - - - - - - - - - - - - - - - - - - -23
      Full workup - - - - - - - - - - - - - - - - - - - - - - - - - - - - -23
        Protein sequences - - - - - - - - - - - - - - - - - - - - - - - - -23
        DNA sequences - - - - - - - - - - - - - - - - - - - - - - - - - -23
    Control of workup levels- - - - - - - - - - - - - - - - - - - - - - - -24
    Security levels - - - - - - - - - - - - - - - - - - - - - - - - - - - -25
    Annotation display - - - - - - - - - - - - - - - - - - - - - - - - - -26
    Show all annotations on all sequences - - - - - - - - - - - - - - - - -26
Contents

Show by category for all sequences - - - - - - - - - - - - - - - - - - 27
Show by Category for one sequence - - - - - - - - - - - - - - - - - - - 28
Show individual annotation - - - - - - - - - - - - - - - - - - - - - - 28
User-created annotations - - - - - - - - - - - - - - - - - - - - - - - - 30
Annotation links - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 30
Edit annotation display - - - - - - - - - - - - - - - - - - - - - - - - - - 31
Data mining - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 33
Tutorial 1: thrombin - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 35
Start with manual search in Look v3 - - - - - - - - - - - - - - - - - - - - 35
Choose the sequence of interest - - - - - - - - - - - - - - - - - - - - - - 36
Load the sequence thrb_mouse - - - - - - - - - - - - - - - - - - - - - - - - 38
Choose workup and privacy levels - - - - - - - - - - - - - - - - - - - - - - 38
Control display of all annotations - - - - - - - - - - - - - - - - - - - - - - 40
Adjust display of entire alignment - - - - - - - - - - - - - - - - - - - - - - 40
Adjust display for one sequence - - - - - - - - - - - - - - - - - - - - - - - 41
Inspect individual annotation classes - - - - - - - - - - - - - - - - - - - - 41
Cross-check annotations - - - - - - - - - - - - - - - - - - - - - - - - - - - - 47
Show annotations for selected positions - - - - - - - - - - - - - - - - - - - - 49
Sequence display attributes - - - - - - - - - - - - - - - - - - - - - - - - - - - 49
Data mining - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 51
Tutorial 2: BRCA1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 52
Start with manual search in Look v3 - - - - - - - - - - - - - - - - - - - - 52
Loading the BRCA1 Sequence - - - - - - - - - - - - - - - - - - - - - - - - - 53
Choose workup and privacy levels - - - - - - - - - - - - - - - - - - - - - - - 54
Load homologous protein sequences - - - - - - - - - - - - - - - - - - - - - - 55
Load a translated ORF into the alignment - - - - - - - - - - - - - - - - - - - - 57

2. Reference Manual: Introduction - - - - - - - - - - - - - - - - - - - 59

About the Reference Manual - - - - - - - - - - - - - - - - - - - - - - - - - 59
Conventions - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 59
Contents

3. Architecture, data flow - - - - - - - - - - - - - - - - - - - - - - - - 61

Chapter contents - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 61
Organization schematics - - - - - - - - - - - - - - - - - - - - - - - - - - - - 62
  GeneMine architecture overview - - - - - - - - - - - - - - - - - - - - - - - 62
  Discovery Engine data flow - - - - - - - - - - - - - - - - - - - - - - - - - - 63
Data flow - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 63
Queries - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 64
  Constraints on queries - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 64
  Services - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 64
  Data manager - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 65
  Security - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 65
  Designation of security level - - - - - - - - - - - - - - - - - - - - - - - - - - - - 65
  Queuing - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 66
  Additive filtering - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 68
  Event-to-service triggers - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 68
WWW- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 68
Virtual documents - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 68
URL formats - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 68
URL checker - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 69
Backup servers - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 69
History stamp - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 69
Caching - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 70
Slow queries - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 70
  Cache hash key - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 71
  External filtering - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 71
  Application recognition - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 71
  Internal filtering/analysis - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 71
Sequence length constraints - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 72
  By GeneMine- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 72
  By servers - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 72
Contents

4. Worldbase - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 75

Chapter contents - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 75
Introduction - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 75
Worldbase organization - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 75
Worldbase format - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 77
Worldbase services - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 77
Worldbase source definitions - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 78
Source - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -78
Service - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -78
History - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -79
Name - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -79
  Example - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -79
Title - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -79
  Example: - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -80
UI - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -80
Path - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -80
Input - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -81
Output - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -81
Maxseq - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -82
And, Or, Not - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -82
Whitespace - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -82
Wildcard - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -82
Strip - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -83
  Example - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -83
  Example - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -83
Viewer - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -83
Filter - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -84
  Example 1 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -85
  Example 2 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -85
Worldbase application definitions - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -85
Application - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -86
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output processing</td>
<td>96</td>
</tr>
<tr>
<td>Homology hit filtered output</td>
<td>97</td>
</tr>
<tr>
<td>Homolog alignment</td>
<td>100</td>
</tr>
<tr>
<td>Homolog clustering</td>
<td>101</td>
</tr>
<tr>
<td>Format rules</td>
<td>103</td>
</tr>
<tr>
<td>Annotation fields</td>
<td>103</td>
</tr>
<tr>
<td>Annotation Types</td>
<td>103</td>
</tr>
<tr>
<td>BLASTHomology</td>
<td>104</td>
</tr>
<tr>
<td>BLASTHomology header formats</td>
<td>104</td>
</tr>
<tr>
<td>BLASTHomology hits format</td>
<td>108</td>
</tr>
<tr>
<td>Example Header</td>
<td>109</td>
</tr>
<tr>
<td>Annotation display</td>
<td>110</td>
</tr>
<tr>
<td>Gap spanning</td>
<td>110</td>
</tr>
<tr>
<td>The 6 calculated annotations</td>
<td>110</td>
</tr>
<tr>
<td>Sequence homolog</td>
<td>110</td>
</tr>
<tr>
<td>50% identity rule</td>
<td>111</td>
</tr>
<tr>
<td>Homolog mapping</td>
<td>112</td>
</tr>
<tr>
<td>Virtual document links</td>
<td>113</td>
</tr>
<tr>
<td>Infolink</td>
<td>113</td>
</tr>
<tr>
<td>Sequence polymorphic regions</td>
<td>114</td>
</tr>
<tr>
<td>Infolink</td>
<td>115</td>
</tr>
<tr>
<td>Structural homologs</td>
<td>116</td>
</tr>
<tr>
<td>Default Display style</td>
<td>116</td>
</tr>
<tr>
<td>SCOP backgrounder</td>
<td>117</td>
</tr>
<tr>
<td>Classification</td>
<td>118</td>
</tr>
<tr>
<td>Infolink</td>
<td>119</td>
</tr>
<tr>
<td>Closely related homolog</td>
<td>119</td>
</tr>
<tr>
<td>Default Display style</td>
<td>120</td>
</tr>
<tr>
<td>EST Expression</td>
<td>120</td>
</tr>
<tr>
<td>EST clusters</td>
<td>121</td>
</tr>
<tr>
<td>Infolink</td>
<td>122</td>
</tr>
</tbody>
</table>
Contents

Homology Fingerprint - - - - - - - - - - - - - - - - - - - - - - - - - - - - 122
Validation of method - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 124
Methanococcus jannaschii genome - - - - - - - - - - - - - - - - - - - - - - - - - 124
Haemophilus influenzae genome - - - - - - - - - - - - - - - - - - - - - - - - - 128

6. Standard annotations - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 131
Chapter contents - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 131
SeqDBAnnotation - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 132
   Document-type: SeqDBAnnotation - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 132
   Motif Annotations - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 133
   Default Display style - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 134
   Secondary structure prediction (SSP) annotations - - - - - - - - - - - - - - - - - - - - - - - - - - - 134
      Example - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 134
      Default display style - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 134
      Accuracy of SSP - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 135
   Domain Annotations - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 136
      PRODOM server - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 136
      Reference - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 136
   Active site - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 136
   Specificity Pocket - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 137
   Epitope - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 137
   Post-translational modifications - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 138
   Metabolic pathway/enzyme activity - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 138
   Genetic map/linkage - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 139
   Disease Association - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 139
   Restriction Sites - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 140
   Open reading frame - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 140
   PCR primer prediction - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 141
   Disulfide bridges - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 144
   Literature - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 144
7. Primary sources - - - - - - - - - - - - - - - - - - - - - - - - - 147

Introduction - - - - - - - - - - - - - - - - - - - - - - - - - - - - 147
Importance of Worldbase - - - - - - - - - - - - - - - - - - - - - 147
Caveat on PATH= variable - - - - - - - - - - - - - - - - - - - - - 148
Primary databases - - - - - - - - - - - - - - - - - - - - - - - - - 148

How Worldbase accesses PDB - - - - - - - - - - - - - - - - - - - 148
Worldbase record for PDB Search at MAG - - - - - - - - - - - - - 148
Worldbase record for PDB at Brookhaven - - - - - - - - - - - - - 149
Worldbase record for PDB at NIH - - - - - - - - - - - - - - - - - 149

Prerelease PDB - - - - - - - - - - - - - - - - - - - - - - - - - - - 150
Worldbase record for Prerelease PDB - - - - - - - - - - - - - - - - 150

SCOP, a derivative of PDB - - - - - - - - - - - - - - - - - - - - - - 150
Worldbase record for SCOP - - - - - - - - - - - - - - - - - - - - - 151

PIR - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 151
Worldbase record for PIR homology alignments - - - - - - - - - - 151
Worldbase record for PIR annotations - - - - - - - - - - - - - - - 152

SwissProt - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 152
How Worldbase accesses Swiss-Prot annotations - - - - - - - - - - 152
Worldbase record for Swiss-Prot Annotations (Main) - - - - - - - - - - 152
Worldbase record for Swiss-Prot Annotations (Backup 1) - - - - - - - - - - - - 153
Worldbase record for Swiss-Prot Annotations (Backup 2) - - - - - - - - - - 153
Worldbase record for Swiss-Prot Annotations (Backup 3) - - - - - - - - - - - - 154

GenBank - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 154
How Worldbase accesses GenBank - - - - - - - - - - - - - - - - - - - 154
Worldbase record for Entrez GenBank - - - - - - - - - - - - - - - - - 154
Worldbase record for GenBank (Alternate Server) - - - - - - - - - - 155
Worldbase record for GenBank (IUBIO) - - - - - - - - - - - - - - - 155
# Contents

8. Collateral sources - - - - - - - - - - - - - - - - - - - 157

**Introduction** - - - - - - - - - - - - - - - - - - - - - - - - - 157

Caveat on PATH= variable - - - - - - - - - - - - - - - - - - - 157

**Combined information sources** - - - - - - - - - - - - - - - - 158

Entrez sources - - - - - - - - - - - - - - - - - - - - - - - - - - 158
GenQuest search (Netscape) - - - - - - - - - - - - - - - - - - - 158
OWL - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 158

**Sequence Databases** - - - - - - - - - - - - - - - - - - - - - - - 159

Worldbase records- - - - - - - - - - - - - - - - - - - - - - 160
Worldbase record for GenQuest search (Netscape) - - - - - 160
Worldbase records for OWL - - - - - - - - - - - - - - - - - - - 160
Worldbase record for sequence databases - - - - - - - - - - 161

**Homology information** - - - - - - - - - - - - - - - - - - - - - - - 161

**BLAST searches** - - - - - - - - - - - - - - - - - - - - - - - - 161
BLAST2 - - - - - - - - - - - - - - - - - - - - - - - - - - - - 162
BLAST flavors - - - - - - - - - - - - - - - - - - - - - - - - - - 162
BEAUTY homology search - - - - - - - - - - - - - - - - - - - 162
GenPept NR BLAST homology - - - - - - - - - - - - - - - - - - 163
GenQuest BLAST homology - - - - - - - - - - - - - - - - - - - 163
SCOP BLAST - - - - - - - - - - - - - - - - - - - - - - - - - - 163

Worldbase records- - - - - - - - - - - - - - - - - - - - - - 163
Worldbase record for BEAUTY - - - - - - - - - - - - - - - - - 163
Worldbase record for BLASTP - - - - - - - - - - - - - - - - - 164
Worldbase record for BLASTX homology annotations - - - - 164
Worldbase record for BLAST (Entrez, SRS links) - - - - - - - 165
Worldbase record for PDB BLAST homology annotations - - - - - 165
Worldbase record for NR BLAST homology - - - - - - - - - - - - 166
Worldbase record for SCOP BLAST homology annotations - - - 166
Contents

Worldbase record for GenQuest BLAST (Swiss) - - 167
Worldbase record for GenQuest BLAST (PDB) - - - 167
BLOCKS Homology Search - - - - - - - - - - - - 168
BLOCKS Homology Search - - - - - - - - - - - - 168
CLUSTAL Sequence Alignment - - - - - - - - - - - - 169
Worldbase record - - - - - - - - - - - - - - - 169
for CLUSTAL - - - - - - - - - - - - - - - 169
dbEST information - - - - - - - - - - - - - - - 169
dbSTS electronic PCR - - - - - - - - - - - - - - - 170
Worldbase records - - - - - - - - - - - - - - - 170
Worldbase record for dbEST transcription - - - - - 170
Worldbase record for dbEST expression - - - - - 170
Worldbase record for dbEST DNA clustering - - - 171
Worldbase record for dbEST protein clustering - - - 172
Worldbase record for dbSTS - - - - - - - - - - - - 172
FASTA homology - - - - - - - - - - - - - - - - - 173
Worldbase records - - - - - - - - - - - - - - - 173
Worldbase record for FASTA NR homology EERIE - 173
Worldbase record for FASTA Swissprot homology EERIE 173
Worldbase record for GQ FASTA homology - - - 174
Feature information - - - - - - - - - - - - - - - - 174
Cloning Vectors - - - - - - - - - - - - - - - - - 174
Worldbase record - - - - - - - - - - - - - - - 175
for Cloning Vectors - - - - - - - - - - - - - - - 175
DNA repeat detection - - - - - - - - - - - - - - - 175
Worldbase record - - - - - - - - - - - - - - - 175
for DNA repeat detection - - - - - - - - - - - - - 175
Domains - - - - - - - - - - - - - - - - - - - - 176
Worldbase record - - - - - - - - - - - - - - - 176
for Prodom annotations - - - - - - - - - - - - - 176
Contents

Eukaryotic Promoters - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 177
MHC peptides - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 177
   Worldbase record - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 177
for MHC peptides - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 177
Open reading frames - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 177
   Worldbase records- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 178
   Worldbase record for ORFs (Any, >25aa) - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 178
   Worldbase record for ORFs (Met, >100aa) - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 178
   Worldbase record for NCBI ORF finder - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 179
PCR primers - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 179
   Worldbase record - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 180
for PCR primers - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 180
PROSITE protein motifs - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 180
   Worldbase records- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 181
   record for Prosite text retrieval at ExPASy - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 181
   for Prosite auto annotations at EBI - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 181
   for Prosite auto annotations at ExPASy - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 182
Restriction sites - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 182
   Worldbase records- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 182
   for Webcutter unique restriction sites - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 182
   for Webcutter (backup) - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 183
Secondary structure - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 183
   Worldbase records- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 184
   for PSSP report server - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 184
   for PSSP prediction server - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 184
Transcription Factors - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 185
   Worldbase record - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 185
for transcription factors - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 185
Translate in 6 Frames - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 185
   Worldbase record - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 186
for Translate in 6 frames - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 186
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature information</td>
<td>186</td>
</tr>
<tr>
<td>BioTechniques</td>
<td>186</td>
</tr>
<tr>
<td>Worldbase record for BioTechniques</td>
<td>187</td>
</tr>
<tr>
<td>ChemFinder Database</td>
<td>187</td>
</tr>
<tr>
<td>Worldbase record for ChemFinder</td>
<td>187</td>
</tr>
<tr>
<td>EC Enzyme Database</td>
<td>188</td>
</tr>
<tr>
<td>Worldbase record for EC Enzyme Database</td>
<td>188</td>
</tr>
<tr>
<td>J. Biol. Chem Full Text</td>
<td>189</td>
</tr>
<tr>
<td>Worldbase record for JBC Online (last 14 months)</td>
<td>189</td>
</tr>
<tr>
<td>MEDLINE References</td>
<td>189</td>
</tr>
<tr>
<td>Worldbase record for MEDLINE</td>
<td>190</td>
</tr>
<tr>
<td>NIH Grants</td>
<td>191</td>
</tr>
<tr>
<td>Worldbase record for NIH grants</td>
<td>191</td>
</tr>
<tr>
<td>Sequence Analysis Refs</td>
<td>191</td>
</tr>
<tr>
<td>Worldbase record for Sequence Analysis Refs</td>
<td>192</td>
</tr>
<tr>
<td>U.S. Patents</td>
<td>192</td>
</tr>
<tr>
<td>Worldbase record for U.S. Patents</td>
<td>193</td>
</tr>
<tr>
<td>Index</td>
<td>195</td>
</tr>
</tbody>
</table>
Contents
1. USER GUIDE: Discovery Engine

USER GUIDE CHAPTER CONTENTS

- Introduction
- Overview: brief discussion of information origins, and annotation formats and categories.
- Discovery Engine Setup: workup levels and security preferences.
- Annotation display: brief description of how to adjust settings.
- Tutorial 1: thrombin: step-by-step example of information discovery based on the keyword search "thrombin."
- Tutorial 2: BRCA1: the same for the keyword search "BRCA1."

INTRODUCTION

ABOUT THE USER GUIDE

The User Guide occupies the first chapter of the combined User Guide + Reference Manual for GeneMine’s Discovery Engine module. The purpose of this User Guide is to describe the Discovery Engine environment in enough detail to let the new user get started, without delving into all the intricacies of the program. For detailed information about Discovery Engine, including changing Worldbase defaults, see the more exhaustive treatment in the accompanying Reference Manual.

The reader may notice a certain degree of redundancy in these documents. This is intentional.
Conventions

Many Discovery Engine operations depend on a knowledge of functionality of the Look v3 core module. Users may find it helpful to refer to the Look v3 User Guide as well as this document while learning Discovery Engine operations.

Elementary operations, such as how to click and drag, how to use a word processor, etc. are not covered in this document.

The colon is used to indicate menu and sub-menu relationships, as in File:New.

Discovery Engine uses the three-button UNIX mouse. Mouse button usage may be referred to by abbreviation. LMB, MMB and RMB mean left, middle and right mouse button respectively.

A working knowledge of UNIX, while not essential, will allow you to manage Discovery Engine files and directories.

Database Licensing

Discovery Engine users should be aware of possible subscription requirements by information providers. The Structural Classification of Proteins database SCOP is one information provider that may require user licensing. Normally academic users are granted free licenses and do not need to register, but commercial users are required to purchase an annual license. A two week free trial period is provided for commercial users. Contact the information provider for details.

SCOP is called during a regular or full protein workup, but not a light workup. If users do not wish to obtain a SCOP license after their 2 week free period, they should disable the scop-blast-annot-auto entry in the Worldbase file.

Discovery Engine Overview

Beginning a Discovery Engine Project

Often the first step in a project is to use a manual sequence or keyword query to search the biological universe. This universe consists of information sources both internal to an organization (intranet) and external public ones (internet).
When a sequence is displayed in the Sequence Window, the next step, automatic sequence annotation, can proceed. The instructions as to where and how the annotation search is performed are stored in "Worldbase" (worldbase.src) a customizable configuration file that contains all necessary information for each of the servers used. The default worldbase.src file is located in the installation directory of Look v3. Users may also have one or more additional copies of Worldbase in personal or site directories; see the Reference Manual for more information on Worldbase.

From the multiple services that can be accessed by Worldbase, sequences, alignments, literature and other information come back. The retrieved information is filtered, sorted and arranged by specifications also resident in Worldbase, and presented at the UI.

**Discovery Engine’s AUTOMATED ANNOTATIONS**

Manual keyword and sequence queries can be performed in the core Look v3 module. Discovery Engine is a separate add-on module that automatically performs multiple sequence-related queries. All automatic annotation retrieval is governed by parameters such as URL source, pattern-matching, etc. which are resident in Worldbase. The initial results are filtered, scored and ranked, and displayed adjacent to the sequences as annotations.

The power and versatility of Discovery Engine are based on three fundamental elements:

- Automation, for truly comprehensive information retrieval. Unlike manual searching, which is bound to be hit-or-miss, Discovery Engine searches biological information sources globally and consistently.

- Filtering, for data reduction and assessment of significance of results. Duplicate results are parsed out, while functional patterns obscured by masses of data are highlighted by clustering. False negatives are rescued, for example by superfamily log-odds scoring. Overlapping data from diverse sources provide multiple independent predictors of reliability.

- Client-side post-processing of large data sets. Where a query returns dozens or hundreds of hits, Discovery Engine analysis functions are applied in a way which would not be practicable on a small, manual scale. For example, in the case of homology "fingerprints," identity patterns across whole protein families are extracted, summarized and displayed on the sequence as annotations.
**Annotations Are Sequence-Associated**

Annotations are only associated with sequences, not with keywords or text. In order to generate automatic annotations, one or more sequences must be present as a query result, or exist already as a sequence file, and be available to load. Typically, it is at the "Show Sequence" stage of a project that creation of automatic annotations is initiated by Discovery Engine, although annotations can be generated at any time through "Sequence Queries."

The extent of annotation searching is controlled by means of "workup levels," which are set prior to launching the request for annotations. "Show Sequence," "Sequence Queries," and workup levels are discussed in detail elsewhere in this document.

Finally, all annotations are hyperlinked to "further information" resources, allowing the user to "drill down" vertically or explore laterally to fully exploit the multiple layers of information that annotations are based on.

**Where Annotations Come From**

There are two broad categories of annotations: calculated annotations and standard annotations. Both types of annotations are generated by queries to information sources such as GenBank, PROSITE, SCOP, ExPASy, etc. These information sources can be categorized in three groups:

- Information obtained from feature table databases. Feature tables are sets of observations and predictions such as disulfide bridge location and domains that are part of sequence records. In many cases feature table data also refer to other databases such as SwissProt, ProDom, OWL, etc. Discovery Engine automatically extracts features from the source databases and displays them as annotations.

- Information obtained directly from servers: some analysis tools such as PROSITE or WebCutter are based on predictive patterns that are applied to the query sequence. Results are displayed as annotations without further processing or secondary searching. Other databases such as Medline simply contain static data. This information, too, is retrieved and displayed as annotations without further processing.

- Information generated by extensive client-side processing of query results: Sequences returned from BLAST and FASTA are first grouped into related families, each of which is then summarized and displayed as a single annotation. This highly processed, calculated type of annotation may apply to entire groups of families as well as to single families.
CALCULATED HOMOLOGY ANNOTATIONS

Discovery Engine calculates homologies in six annotation categories:

- Sequence Homology
- Sequence Polymorphisms
- Structural Homology
- Close Homology
- EST Expression
- Homology Fingerprint.

When given a new query sequence, Discovery Engine searches for homologs using a variety of sources, including SCOP, GenBank, SwissProt and dbEST. (This version of Discovery Engine uses BLAST and FASTA to search these sources. Installations can be customized to use Smith-Waterman or other processes as well; see the Reference Manual). The outputs from those sources are standardized using filters, so that all results can be interpreted in the same way. Finally, Discovery Engine clusters the significant homologs into families and calculates annotations based on these families.

ANNOTATION SCORING & UNCERTAINTY

Many annotations are feature predictions based on sequence analysis. As such, they have inherent uncertainty, typically quantified as a probability value (p-value) which estimates the likelihood that the prediction was obtained by random chance (i.e., null hypothesis).

To the extent possible, Discovery Engine seeks to derive a probability score estimating the level the level of uncertainty for each annotation as a log odds (lod) score. The lod score, the negative logarithm of the p-value, has several advantages because it is easy to represent, interpret, and calculate, and can be summed for separate predictions to estimate their joint probability. For example, a lod score of 1 corresponds to a $10^{-1}$ (10%) probability that the predictor could have been obtained by random chance.

In Discovery Engine a user-adjustable preference sets a threshold score below which annotations are marked as "uncertain." By default this threshold value is 1.5 (p-value of 3%). Uncertain annotations are indicated by italics.
**Discovery Engine User Guide**

**HOMOLOG CLUSTERS**

In order to reduce the amount of information returned by homology searches, Discovery Engine groups the homologs into families, or "clusters." "Sequence Homology" and "EST Expression" families are based on sequence identity. Structural clustering is based on SCOP protein superfamily classification system developed by Alexei Murzin and his colleagues. Discovery Engine groups homologs into families by a 50% rule: any sequence with greater than 50% identity to the scoring sequence of a family joins that family. This family-clustering threshold is user-adjustable.

Sequence identity is defined as the percentage of identical amino acids in the alignment between two sequences. At least 20 amino acids must match, regardless of sequence length, for sequence identity to be calculated; sequences with fewer amino acids will be reported as having 0% identity. The maximum number of families allowed is 100.

**STANDARD ANNOTATIONS**

Standard annotations are derived from feature tables of source databases. Aside from output format standardization, no further processing is done on them at the client side.

There are sixteen standard annotation categories in this version of Discovery Engine:

- Domain
- Active site
- Protein secondary structure prediction
- Specificity pocket
- Motifs
- Post-translational modifications
- Numbering
- Metabolic pathway/enzyme activity
- Genetic map/linkage
- Disease association
- Restriction sites


**Discovery Engine User Guide**

- Open reading frames
- PCR primer generation
- Disulfide bonds
- Literature

**Discovery Engine ANNOTATIONS**

This section briefly describes the structural and sequence features represented by each annotation type, and its default display style.

1. **DOMAIN**

Region (range of sequence positions) known or predicted to represent a functional domain or region. Domain annotations may be taken from feature table databases (e.g., SwissProt, GenBank, PIR), domain predictions (e.g., PRODOM), or other predictions such as transmembrane segment.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

**Examples**

Signal peptide, pro-peptide, ligand-binding domain, trypsin-sensitive domain, transmembrane region.

**Default display style**

Solid white line spans domain region; text identifier (e.g., PROTEIN C HEAVY CHAIN domain).

2. **ACTIVE SITE**

Sequence (residue) positions in enzymes which are known or predicted to be involved in catalytic activity of a protein. Taken from feature table databases (SwissProt).

Click "About this annotation" to view statistical and other supporting data in the Information Window.
Example

Serine protease catalytic triad (His/Asp/Ser).

Default display style

Red * at active site position(s); text identifier (e.g., ACT_SITE CHARGE RELAY SYSTEM).

3. SECONDARY STRUCTURE

Helix, sheet, turn, coil structures. In PDB files this information, if present, is based on observation by the author of the record. For non-PDB files, secondary structure is predicted. Prediction is applied to all protein sequences, regardless of whether PDB secondary structure is present, so that PDB files may display two sets of secondary structure annotations.

Click "About this annotation" to view statistical and other supporting data in the Information window.

Example

Helix, turn, sheet, or random coil (i.e., no secondary structure).

Default display style

<table>
<thead>
<tr>
<th>SwissProt, etc.: Predicted</th>
<th>Helix</th>
<th>Strand (Sheet)</th>
<th>Turn</th>
<th>Random Coil</th>
<th>Type style</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Open box</td>
<td>Open arrow indicating N-to-C directionality</td>
<td>&quot;=&quot; symbol (SwissProt only) above each residue</td>
<td>—</td>
<td><em>italics</em>, to indicate uncertainty</td>
</tr>
<tr>
<td>PDB: Observed, recorded by author</td>
<td>Solid cylinder with black dot at N-terminal end</td>
<td>Solid arrow indicating N-to-C directionality</td>
<td>&quot;</td>
<td>&quot; symbol above each turn residue</td>
<td>solid line</td>
</tr>
</tbody>
</table>

Table 1. Secondary structure display styles
Each sequence in an alignment has its own color for secondary structure symbols. Colors are allocated from white to red, in the order in which they appear in the color palette.

For PDB records which do not contain secondary structure information, a solid white line running the length of the sequence is displayed.

4. **SEQUENCE HOMOLOGS** *

Sequence homology annotations are obtained through these steps:

- Sequence similarity search by BLAST/FASTA.
- Low-stringency filtering of individual hits for significance.
- Reduce data bulk by clustering into homology families, using structural superfAMILY classification, when available, or the 50% identity rule.

From these clusters, both "Most Conserved" sites and "Most Deleted" (see "Sequence Polymorphisms") are calculated. A Most Conserved region is a highly conserved sequence region that is likely to be associated with a functional element. A Most Deleted region is a non-conserved region of sequence that is likely to be associated with a loop at the surface of the protein. See also "Homology Fingerprint" and "Sequence Polymorphisms."

Click "About this annotation" to view statistical and other supporting data in the Information Window.

**Default display style**

Homologies are indicated by a single line for each family. Families comprised of a single sequence, or multiple sequences that have very close sequence identity (by default 95%) are shown with a turquoise line. Families consisting of more diverse sets of sequences are shown with a white line.

Text adjacent to line identifies homolog, percent identity, and (in parentheses) number of proteins in homologous cluster (e.g., 39% identity (8): DNA POLYMERASE pir).

5. **SEQUENCE POLYMORPHIC REGIONS** *

Sequence positions where deletions or mutations are observed are annotated. Deletions in close homologs indicate the likely positions of putative loops.
**Discovery Engine User Guide**

This annotation uses the alignment of all homologs to the query sequence. At each residue position of the query sequence, a deletion score is calculated, the percent identity of the closest homologs which have a deletion at that position. Discovery Engine marks the 10-25% of the residues that have the highest deletion scores as **Most Deleted**.

Sequence polymorphisms are the complement of the "Most Conserved " super-annotation. That is, they show the highest-scoring differences between the query sequence and the homolog clusters.

This annotation category also displays mutation data on known polymorphisms and their biological effects, where available from feature table databases.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

**Default Display style**

Cyan ✗ character and text identifier for Most Deleted (e.g., *Most Deleted dds of Genbank NR protein homologs*), italicized to indicate uncertainty.

Blue ▲ character to indicate known mutations (e.g., M -> T (IN LFS))

6. **Specificity Pocket**

Include binding sites and prosthetic groups, from feature table database information in SwissProt.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

**Examples**

Heme, flavin, etc.

**Default display style**

Orange ▲ character at site, with text identifier (e.g., **BINDING HEME (CO-VALENT))**.

7. **Epitope**

(Place-holder in this version of Discovery Engine).
8. **FUNCTIONAL MOTIFS**

Sequence patterns that match characteristic protein features are often not detected by sequence similarity searches. Functional motifs are predictive, being based on matches between the query sequence and the PROSITE database, or other pattern search methods such as PFAM, to identify potential functional features. Common matches that are likely to be false positives, such as N-glycosylation, found in most known protein sequences, are omitted.

Click "About this annotation" to view statistical and other supporting data in the Information window.

**Default Display style**
Green line spanning motif region, with text identifier (e.g., motif Scorpion short toxins signature).

9. **STRUCTURAL HOMOLOGS**

Sequences found by sequence similarity search (see "Sequence Homologs"), by a BLAST search against the SCOP database. Sequence hits are clustered most typically by the SCOP superfamily classification system, so that sequences belonging to the same broad fold and functional class are clustered together. These hits may provide templates for modeling query sequences of unknown structure. SCOP is the "Structural Classification of Proteins" — a database and server in Cambridge, U.K. Click "About this annotation" to view statistical and other supporting data in the Information Window.

Note that homolog clusters of sequences that contain a PDB hit are classified under Structural Homologs.

**Default Display style**
Double line spanning region of homology, with text identifier of homolog: the structural superfamily name for the sequences in the cluster. Percent identity and number of proteins in the SCOP superfamily are indicated (e.g., 44% identity (11) coagulation factor IX).

Families comprised of a single sequence, or multiple sequences that have very close sequence identity (by default 95%), are shown with cyan lines. Families consisting of more diverse sets of sequences are shown with white lines.

10. **POST-TRANSLATIONAL MODIFICATIONS**

Many proteins have amino acids which are modified chemically after the protein is synthesized. This information, when known, is sometimes avail-
Discovery Engine User Guide

able in the feature table database. Discovery Engine extracts this information and shows it graphically against the sequence as for other annotations. Click "About this annotation" to view statistical and other supporting data in the Information Window.

Examples

Glycosylation (N- or O-), amidation, hydroxyproline.

Default display style

Green character at modification site, with text identifier (e.g., MOD_RES HYDROXYLATION).

11. NUMBERING

Sequence position as retrieved from the feature table database (PIR/SwissProt/NBIR/GenBank etc.).

Default display style

Sequence position numbers, by tens, in white.

12. CLOSELY RELATED HOMOLOGS*

Generated by BLAST and FASTA searches against many databases, the same kind of cluster and family alignment — pairwise alignments and grouping into sub groups and families — is carried out with these annotation alignments as with sequence alignments. This annotation class represents sequences with very close identity to the query sequence. By default, the minimum percent identity required is 90% for the top cluster member.

Click "About this annotation" to view a statistics table for the cluster in the Information window; follow a blue link to view the raw FASTA/BLAST results.

Default Display style

Graphic: solid white bar, with text identifier of close homolog, including percent identity and number of proteins in closely homologous group (e.g., 71% identity (17): PROTEIN C PRECURSOR).
13. EST EXPRESSION*

This annotation is calculated in a way similar to the Sequence Homology and Structural Homology annotations: a BLAST search is done databases such as dbEST, and the families are clustered according to sequence similarity. Additionally, a dictionary of expression terms (e.g., "heart" "brain" "thymus" etc.) is searched against all the results, to elucidate patterns of tissue expression from the ESTs. The expression dictionary (expression.hot) is located in the Look/Discovery Engine install directory and is editable at the UNIX shell. The utility of this list is to extract and display a meaningful term, if present, from the EST file header information. The EST Expression dictionary is user-customizable.

Discovery Engine has two clustering schemes for EST data:

- Clustering into gene families, with a 50% identity cut-off, the same default value as in other homology searches.

- Clustering by individual genes, with a 95% identity cut-off. This option can be made the default by editing the Worldbase configuration file, or can be applied directly by the user via the Sequence Queries dialog, as shown below.

Access EST annotations from Sequence:Sequence Queries. Among the many available options in the pull-down are "dbEST Expression Annotations" for protein sequences, and "dbEST Transcription Annotations" for DNA sequences. The default cut-off for families is 95% identity.

Click "About this annotation" to view statistical and other supporting data in the Information Window.
Discovery Engine User Guide

14. METABOLIC PATHWAY/ENZYME ACTIVITY

When an EC (Enzyme Commission) number is present in the record header, metabolic pathway information is retrieved from the ExPASy ENZYME Database.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

Example

Lysyl endopeptidase, from Achromobacter lyticus.

Default display style

White line with text identifier (e.g., EC Classification 3.4.21.50)

15. GENETIC MAP/LINKAGE

A range of sequence positions matched against an STS. Connects sequence to chromosomal maps - done by sequence similarity searches against the STS database. Similar to ESTs, STSs are "Sequence-Tagged Sites" - small regions of genomic DNA that have been cloned and sequenced from a known chromosomal location.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

Default display style

Blue line spans match region, accompanied by text identifier (e.g., STS 21899 on Homo sapiens Chromosome 2).

16. DISEASE ASSOCIATION

Disease association terms (e.g., carcinoma, diabet, etc.) are searched against all the results to elucidate patterns of expression from the ESTs. The dis-
ease association dictionary (diseases.hot) is located in the Look/Discovery Engine install directory and is editable at the UNIX shell to allow customization of this dictionary by the user.

The disease association dictionary is cross-category and could include hits categorized as Domain, Homology and Motif annotations. Disease associations will be displayed when any of the relevant annotation categories are toggled on. Conversely, when Disease annotation is turned on, all other annotation categories (and their icons) in which the disease term is found will also be turned on.

Click "About this annotation" to view STS record showing experimental source, procedure, detailed mapping information, local homologies, and other supporting data in the Information Window.

**Default display style**

As for the annotation type whence the disease association originated (Domain, Homology, Motif, etc.).

**17. HOMOLOGY FINGERPRINT**

Following BLAST and FASTA searches against many databases, patterns of sequence identities and deletions are tabulated across the total set of sequence homologs. These patterns can be displayed with the sequence homolog cluster "lines" ("Sequence Homologs") or by themselves to highlight key functional regions across sequence families.

This category of annotations can be split into Most Conserved and Family Homology.

**Most Conserved**

The Most Conserved annotation marks the residues that are most conserved for all the homologs found by each type of search (e.g., SCOP, SwissProt, dbEST) for the query sequence. A conservation score is calculated for each residue position of the query sequence. This score is the maximum percent identity of all homologs for which the amino acid at that position is not identical to the one in the query sequence. The top 10-25% of the residues with the highest conservation scores is marked as core homologies, the "fingerprint residues."

Being based on values calculated by Discovery Engine, annotation links through the "About This Annotation..." command are not available for core homologies.
**Default display style, Most Conserved**

Cyan ▲ character for each conserved residue, in linear arrays. Each line represents a homology cluster.

**Family Homology**

The Family Homology annotation is similar to the Most Conserved annotation, except that instead of being calculated for all homologs, it is calculated for each family. Because there are fewer sequences in the calculation, the statistics are coarser, and the top 10-50% of the residues with the highest conservation scores is marked.

This annotation is only calculated if the sequence identity between the query sequence and the family is less than a default threshold of 70%.

**Default display style, family homology**

Purple ▲ character for each conserved residue, in linear arrays. Each line represents a homology cluster.

18. **RESTRICTION SITES**

(DNA annotation only) Shows unique restriction sites, indicating enzyme recognition sites that occur only once in the DNA segment. Performed using the "WebCutter" program through the Baylor College of Medicine site. The range of nucleotides constituting the recognition site is shown.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

**Default display style**

Bracketed white line at cut site, with name of enzyme (e.g., HindIII).

19. **OPEN READING FRAME**

(DNA annotation only). Prediction of regions in DNA which might be translated into protein sequences. For cDNA sequences, these should be contiguous. For genomic DNA, introns of varying size (length) and number may be present between the exons, complicating the analysis of potential open reading frames. Discovery Engine by default uses the NCBI "GORF" server for this category.
Discovery Engine also performs a six-frame translation search against protein sequence databases to discover potential open reading frames by homology. Results of this search, including links to load the inferred amino acid translation, are shown as homology annotations.

Click "About this annotation" to view in the Information Window the translation, and the link to load it for workup, alignment and analysis.

**Default display style**

White line with arrowhead to show directionality of ORF. Text label indicates frame number.

**20. PCR PRIMER PREDICTION**

(DNA annotation only). Displays ranges of sequences which might be useful for creating synthetic DNA primers for PCR to clone out DNA fragments from larger stretches of DNA. Displays location and orientation (arrow) of the predicted possible primer sequences. Obtained by sending sequence to Baylor College of Medicine site which uses the xprimer program.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

**Default display style**

Inward-pointing arrows at forward and backward primer sites.

**21. DISULFIDE BRIDGES**

Often a key structural element from feature table databases.

Click "About this annotation" to view information extracted from the feature table of the sequence record in the Information Window.

**Default display style**

Yellow dot below cysteine residue, with text label "S-S bridge," numbered in the order of occurrence in the source record’s Feature Table.

**22. ANNOTATIONS & LITERATURE**

List all annotations and literature citations in the current session. This list comprehensively records all pointers that are accessible individually from
"About this Annotation." Typically the bibliographic links lead to Entrez Pubmed, but may also lead to OMIM and other databases.

A literature icon appears in the Sequence Window annotation cluster if the sequence record (in the Information Window) contains links to Medline. This icon is limited to papers and does not include annotation links. Note that neither of these icons appears on the Annotation Toolbar.

Default display style

White line (normally spanning the entire sequence) with bibliographic citation.

**ANNOTATION FORMAT & DISPLAY**

Annotations are formatted at the source differently according to the server and the type of information. Formatting is recognized by means of scripts called by Worldbase. After processing, annotations are available to be selected and displayed below the pertinent sequence position or range in the Sequence Window. Use the annotation icons in the Annotation Toolbar to display annotations globally. Use the annotation icons at the left side of the Sequence Window to display annotations for particular sequences.

**USER-DEFINED SEARCH TERMS**

Discovery Engine contains three user-editable lists of terms at the user level, which it uses to scan all incoming information.

**HOTLIST ANNOTATIONS**

Hotlist terms are searched for in all returned information, regardless of its relevance to any particular search. Hits to any of the "hotlist" terms are flagged hot red, and Discovery Engine beeps. Hotlist annotations override annotation icon toggles and are displayed even if icons are turned off.

The hotlist file is called hotlist.hot and is located in the directory .look within the user's personal account. It is editable at the UI. To add or remove terms from the hotlist, go to Edit:Preferences:Annotation and click on the "Hotlist Terms" button. The hotlist form will appear in the Information Window. To add a non-searching comment line to the hotlist, start each comment line with a bang (!).
DISEASE ASSOCIATION ANNOTATIONS

A second dictionary (diseases.hot) is kept for disease associations. Unlike the hotlist, which is personal, the disease association dictionary is global. It resides in the Look/Discovery Engine install directory. It is not editable at the UI. Site administrators or users who have the proper permissions can edit the diseases.hot file by going to the install directory at the UNIX shell and opening the file with a text editor such as emacs or vi.

The disease association dictionary is cross-category and could include hits categorized as Domain, Homology and Motif annotations. Disease associations will be displayed when any of the relevant annotation categories are toggled on.

EXPRESSION ANNOTATIONS

A third global dictionary (expression.hot) for identifying expression patterns from EST hits also resides within the Look/Discovery Engine install directory. It contains names of tissues, such as marrow, lung, liver or brain. The utility of this list is to extract and display a meaningful term, if present, from the EST file header information. Otherwise the annotation would be identified only by its accession number.

Like the Disease Association list, the Expression list is editable at the UNIX shell. Site administrators or users with the proper permissions can also customize this system-wide dictionary.
**Discovery Engine User Guide**

**USER-CREATED ANNOTATIONS**

Users can easily create their own annotations:

1. Select the desired sequence positions representing a feature within a single sequence.
2. Go to Sequence:Add Annotation. Be sure to give the annotation a name.
3. Adjust styles as desired

![Figure 3. Edit annotation dialog](image)

4. Any document in the Information Window browser can be linked to an annotation (see the discussion in the Look v3 + Modeling Package User Guide).

**Discovery Engine SETUP**

**WORKUP LEVELS**

Annotations are acquired when the downloaded sequence is transferred from memory to display. At this important moment you are prompted to set the extent
of annotation to be applied to the sequence by choosing one of four "workup levels".

![Workup level prompt](image1)

**Figure 4. Workup level prompt**

Annotations can also be acquired for a sequence already in an alignment, or a sequence can be re-annotated, with commands in a similar pulldown in Sequences:Sequence Queries.

![Sequence Queries](image2)

**Figure 5. Workup level + servers, in Sequence Queries**
Control over workup level and data servers is also available in the Sequence Window by pressing the button and then clicking on "Search/Analysis.

The definition of specific annotation types for each of the workup levels is configurable in the worldbase.src file. Default workup levels are described in the following sections.

**NO WORKUP**
Numbering of sequence only

**LIGHT WORKUP**

**PROTEIN SEQUENCES**
Secondary Structure Prediction.
Prosite Motif Detection.

**DNA SEQUENCES**
Numbering of sequence.
NCBI ORF Finder.
DNA Repeat Detection.

Regular workup
Discovery Engine User Guide

PROTEIN SEQUENCES
Secondary Structure Prediction.
Prosite Motif Detection.
SCOP BLAST Homology Annotations.
FASTA PIR Homology Annotations EERIE (about 60,000 seqs).
dbEST Expression Annotations.

DNA SEQUENCES
Numbering of sequence.
NCBI ORF Finder.
DNA Repeat Detection.
BLASTX Homology Annotations.
dbEST Transcription Annotations.

FULL WORKUP

PROTEIN SEQUENCES
Secondary Structure Prediction.
Prosite Motif Detection.
SCOP BLAST Homology Annotations.
FASTA NR Homology Annotations EERIE (about 220,000 seqs).
dbEST Expression Annotations.

DNA SEQUENCES
Numbering of sequence.
NCBI ORF Finder.
DNA Repeat Detection.
BLASTX Homology Annotations.
dbEST Transcription Annotations.
dbSTS Electronic PCR.
Discovery Engine User Guide

Webcutter Unique Restriction Sites.
PCR Primer Generation.

CONTROL OF WORKUP LEVELS

Workup levels for individual sequences can be set using any of the three dialogs illustrated in Figs. 4-6.

Workup levels can be adjusted globally in the Edit:Preferences:Annotations dialog.

![Annotation Preferences dialog](image)

Figure 7. Workup levels, global control

Note that this dialog allows separate control of annotations of new sequences and close homologs. The rationale for separate control of close homologs is that a sequence closely related to one in the alignment which has already been annotated may be redundant in terms of information it is likely to contribute. The homologous sequence therefore may not require as extensive a workup as a new, unknown sequence. Accordingly, the user can set two separate preferences establishing the default workup level for these two classes of annotations.

By default, Discovery Engine is set to Regular Workup for new sequences, and Light Workup for close homologs.
Display settings can also be set independently for new sequences and close homologs. For example, it might be desirable to display fingerprint annotations for new sequences but not for close homologs, and to show functional motifs for both.

**SECURITY LEVELS**

To avoid electronic eavesdropping, sequences sent out on net annotation searches are designated by security level. Network server security levels are configurable in the worldbase.src file. There are four security levels: public, regular, strict, absolute. The local site administrator sets site policy for what kinds of queries are allowed at each level. The user can set the privacy level for individual sequences at the same time the workup level is determined.

This Worldbase-defined security layer ensures that no proprietary data can be sent to a source not meeting the security threshold for the specific data. Information about blocked queries is presented at the title bar.

Note that the user must choose the security level for each sequence loaded. This system can only provide protection against accidental sequence submission, but cannot prevent intentional release of proprietary sequences to public servers.
ANNOTATION DISPLAY

After a search is launched, information is automatically retrieved, processed and displayed as annotations. How annotations are displayed can be controlled by the user at several levels, from all annotations of all sequences down to single annotations.

SHOW ALL ANNOTATIONS ON ALL SEQUENCES

The View:Toggle sub-menu contains the "Hide All Annotations" toggle. With this toggle unchecked, all annotations retrieved in the search are displayed next to the sequence in the Sequence Window. Toggle on "Hide All Annotations" to clear the annotation display. When toggled off again, annotations may be selectively displayed by using the icons in the Annotation Toolbar or the Sequence Window. Note that the "Hide All Annotations" command does not remove annotations from memory, but merely temporarily suppresses their display.

Figure 9. All annotations displayed
SHOW BY CATEGORY FOR ALL SEQUENCES

To display annotations by category for all sequences, turn on the icons in the Annotation Toolbar. Display the Annotation Toolbar with the View:Annotation Toolbar... toggle.

Icons representing currently displayed annotation classes are shown in red. Icons standing for annotation classes that are available but not displayed are shown in grey with black writing. Annotation classes for which no annotation is available (because the workup level or server did not include that category) are shown in grey with white writing.

Note that with multiple annotation categories displayed, the Sequence Window may become filled with so much information that the image may become overcrowded. Accordingly, it’s wise to display annotations selectively.
SHOW BY CATEGORY FOR ONE SEQUENCE

To display annotations by category for an individual sequence, use the icons at the left of the Sequence Window. These buttons control categorical annotation display at the sequence level. The image shows control buttons for six annotation classes for two sequences.

Figure 12. Annotations by sequence

This control region to the left of the sequences also contains:

- Sequence name
- Colored dot representing family clustering
- Individual annotation icons (red=displayed, gray=not displayed)
- button for controlling alignment and annotation display.

SHOW INDIVIDUAL ANNOTATION

To pick out a particular annotation to display, Use the Styles:Annotation Styles command.
Figure 13. Individual annotations listed

Pick the annotation(s) of interest, then click OK. The Set Annotation Styles dialog comes up. Use the Show and Hide buttons to show or suppress display of annotations selected in the previous step. Also use this dialog to set the display attributes of the annotation of interest.

Figure 14. Set Annotation Styles dialog

Click Done. The annotation will appear next to the sequence.

Figure 15. Individual annotations displayed
USER-CREATED ANNOTATIONS

To add your own information in form of an annotation, first select the sequence position of interest. Use the Pick This dialog. (For a review of Picking procedure see the discussion in the Look v3 + Modeling Package User Guide).

Next, go to the Sequence menu and click the Add Annotation command. The Edit Annotation dialog comes up. It contains an assortment of various display options, including positioning, color, disposition of gaps, and icon associations.

![Edit annotation dialog](image)

Figure 16. Edit annotation dialog

Make sure to select "Show Annotation", "Show Label" and a symbol including its color. When you are done, click "OK". The new annotation will appear in the display in the Sequence Window.

ANNOTATION LINKS

To link an annotation to a user-created hypernote or a hyperlink such as a Medline reference, click on the button at the top of the Information Window to bring up the Snapshot drop-down list.
Click on "Link to annotation..." A dialog appears. Find the relevant annotation in the list and click "Link." The link will be established.

In the future when you select "About this Annotation" in the Annotation po-up menu, the linked document will appear in the Information Window.

**EDIT ANNOTATION DISPLAY**

To change the appearance of one or more annotations, go to Styles:Annotation Styles... A dialog appears.
Discovery Engine User Guide

Select one or more annotations from the list and click "OK". A second dialog appears.

Make the desired adjustments and click "Done." The display will be altered accordingly.

To change the appearance of a particular annotation, hold down the mouse on that annotation in the display. The Annotation Pop-up menu appears.
Select "Edit This Annotation" from the pop-up menu. The Edit Annotation dialog comes up. Make changes to the annotation as described in "User-created annotations," above.

**DATA MINING**

To get more detail on any annotation, go to the Sequence Window, place the cursor over the annotation and hold down the left mouse button. The following dialog will appear.

Click on "About this Annotation" and hyperlinked information about the Annotation will load to the Information Window. This information can include search results for the Annotation, explanatory notes, and links to related molecules, sequences and other information sources. The linked information is specified by the INFOLINK= statement in the Worldbase file.
In the case of a family homology, a report is generated on the fly and displayed in the Information Window.

![Family breast/oVarian cancer susceptibility protein](image)

**Figure 23. Annotation report**

Follow these links to "drill down" into the information on which this annotation is based. By following the links, important related information can be identified and incorporated into the project.
TUTORIAL 1: THROMBIN

START WITH MANUAL SEARCH IN Look v3

To explore Discovery Engine’s automated annotation capabilities, in this protein-based example we start by searching "Sequence Databases" with the keywords "thrombin and mouse".

We will locate a protein sequence, download it, study its annotations as they are displayed, then "drill down" to discover more information connected to individual annotations.

Go to the menu File:New:Keyword Searches and select "Sequence Databases," which is a collection of several servers searchable by keyword. Type "thrombin and mouse" and click on "Search" to launch the search.

Progress is indicated on the title bar. To check on progress of a search, click on the Inbox icon of the main tool bar and view the ongoing searches in the information box that comes up.

The search result is returned to the Information Window.
CHOOSE THE SEQUENCE OF INTEREST

Search results ("hits") matching the query are returned to the Information Window as an alphanumerically sorted list. Each entry (record) has a link, indicated by red highlighting, to the source database. Scroll through the list until you come to "swiss:THRB_mouse PROTHROMBIN PRECURSOR (EC 3.4.21.5)".

To bring the record itself into the Information Window, hold on the link to display a context-sensitive menu where one or more links may be displayed. Bypass the menu and load the sequence directly by simply clicking on the red list link. The sequence record will appear in the Information Window.
Figure 26. Sequence record

Click on the left "back" arrow above the Information Window to return to the search result.
LOAD THE SEQUENCE THRB_MOUSE

Hold down the mouse on the same record and a context-sensitive menu will appear. Click on "Load Sequence...". The sequence will load. Progress is indicated on the top title bar. After the Sequence downloads to Discovery Engine memory the "Choose Sequence(s) to Show" dialog comes up. The dialog list shows the sequence name.

CHOOSE WORKUP AND PRIVACY LEVELS

The process of displaying the sequence in the Sequence Window will initiate multiple searches for many different types of information. Before proceeding with this crucial step, it's important to set the workup level and privacy controls.

In the "Choose Sequences to Show" dialog, indicate workup level (i.e., extent of Annotations to be obtained). For this example choose "Full." Also indicate privacy level, and whether the sequence is to be added to a new or existing alignment. You're also given a choice of whether to align related DNA and protein sequences in the same alignment, in cases where both DNA and protein sequences are available.

When done, select the sequence and click on "OK". The not-yet annotated sequence appears first in the Sequence Window. Discovery Engine will immediately start annotating the sequence. Progress is indicated in the title bar.
Annotations will appear rapidly and are displayed below their associated sequence positions in the Sequence Window.

Note that annotations can also be acquired for a sequence already in an alignment, or a sequence can be re-annotated, with commands in Sequences:Sequence Queries... The procedure differs from that of a new sequence in that there is no option for setting privacy level. The privacy level set during the initial search also applies to subsequent searches, unless they are changed. Privacy level settings can be changed using Sequence:Edit Sequence at the main menu, or in the pop-up menu in the Sequence Window.

Otherwise, the procedure is the same: first select the sequence, then the workup level, then click OK. The search will be launched and the annotations returned to the Sequence Window.
**Discovery Engine User Guide**

Figure 29. Sequence Queries dialog, for re-annotating

**CONTROL DISPLAY OF ALL ANNOTATIONS**

Display of all annotations for all sequences can be toggled in the menu View:Toggle:Hide all Annotations. Note that the "Hide All Annotations" command does not remove annotations from memory, but only temporarily suppresses their display.

**ADJUST DISPLAY OF ENTIRE ALIGNMENT**

With all annotations on, the display is apt to be too crowded and confusing.

To control display of the individual annotation classes for all sequences in the alignment, go to the menu View:Annotation Toolbar and click on the annotation icons.
Currently displayed annotation classes are shown in red. Available but not displayed annotation classes are shown in grey with black writing. Annotation classes for which no annotation is available are shown in grey with white writing.

**ADJUST DISPLAY FOR ONE SEQUENCE**

To turn annotations on or off by category for an individual sequence, use the icons to the left of the sequence display.

Next inspect the sequence from various perspectives, using the available annotation categories. Note that for categories related to sequence homologies, Discovery Engine is in fact doing considerable client-side processing to produce the final annotations from the original BLAST/FASTA outputs returned by the servers.

Control the display by clicking on the icons either next to the sequence for by-sequence display, or in the Annotation Toolbar for global display.
**Discovery Engine User Guide**

Display Domains.

![Image of domain annotations]

Figure 32. Domain annotations

Display Active Sites.

![Image of active site annotations]

Figure 33. Active site annotations

Display Secondary Structure Prediction.

![Image of secondary structure annotations]

Figure 34. Secondary structure annotations
Display Sequence Homologs. White lines indicate families with multiple homologous sequences, cyan lines indicate a single or multiple very closely related (almost identical) sequences.

![Homology annotations](image)

**Figure 35. Homology annotations**

Display Polymorphic regions. These annotations are based on homology analysis and indicate where the sequences and families and your query sequence are especially divergent and likely to contain loop regions.

![Most Deleted annotations](image)

**Figure 36. Most Deleted annotations**

Display Functional Motif (directly obtained from Prosite)
**Discovery Engine User Guide**

Figure 37. Functional motif (Prosite) annotations

Display Structural Homology, found in SCOP database (Structural Classification Of Proteins).

Figure 38. Structural homology (SCOP) annotations

Display post-translational modifications (found in the SwissProt feature table database).
Discover Engine User Guide

Figure 39. Post-translational (SwissProt) annotations

Display Numbering and Miscellaneous.

Figure 40. Sequence numbering annotation

Display closely related homologs; white dashed line indicates 90% or greater similarity score.

Figure 41. Closely related homolog annotations

Display homology to expressed sequence tags (ESTs).
Discovery Engine User Guide

Figure 42. EST homology annotations
Display Homology Fingerprints. Purple cross indicates homology between sequence and one protein family; cyan cross indicates Most Conserved.

Figure 43. Homology fingerprint annotations
Finally, display disulfide bonding.

Figure 44. Disulfide bridge annotations
CROSS-CHECK ANNOTATIONS

We can estimate by consensus the validity of the annotations by checking to see if the same annotation is reported from different sources. For example, we can compare structural homology and disulfide bonding annotations. In this case, we find that disulfide bridges in the Prothrombin sequence are found in the same regions as structural homologies. Since these information sets are derived from separate sources, the fact that they correspond in this instance tends to give independent confirmation of the accuracy of the information.

![Figure 45. Correspondence between annotation classes](image)

We follow the same approach by comparing a Kringle Domain motif with annotations indicating homology to structures also containing Kringle motifs.

![Figure 46. Kringle from Prosite](image)
In a last step, we add global identities by looking at core (cyan cross) and family identities (purple cross) in Homology fingerprint annotations.

- **Purple cross**: Homology between your sequence and members of one protein family.
- **Cyan cross (core)**: Identity between your sequence and all identified protein families (one set for each of 3 possible types: EST, SCOP, NR).

It becomes evident that a Kringle domain is indicated independently as a Prosite motif and as a homology to Kringle domain-containing structures. At the same time, the same region shows high homology between the Prothrombin sequence...
and the identified sequence families. This indicates that Kringle domains are important in several proteins, and not only in Prothrombin.

**SHOW ANNOTATIONS FOR SELECTED POSITIONS**

To get more information about annotations relevant to an entire sequence or selected sequence positions, click on the icon at the Main Toolbar. The "Click Papers to View" dialog comes up with a list of annotations. Each annotation refers to the related information for that annotation, and is equivalent to clicking on "About this Annotation" in the alignment.

![Figure 49. View papers about annotations](image)

To get more information on one of the annotations, click on the one of interest, then click Done. The results are returned to the Information Window and allow you to follow the links to discover more about the selected annotations.

**SEQUENCE DISPLAY ATTRIBUTES**

Adjust the display to accentuate sequence and alignment features of interest. The human visual system is good at picking out visual cues, and GeneMine makes use of this capability. Together with the annotations provided by Discovery Engine, these cues can add valuable contributions to your analysis.

The menu View:Highlight... shows a drop-down menu offering many options for highlighting a sequence or an entire alignment according to various color schemes.
Display attributes can be adjusted to show:

- Hydrophobic/hydrophilic residues
- Charged residues
- All cysteines
- Beta strands
- Steric
- 4, 5, or 7 amino acid types
- Function
- turn
- Helicity
- By annotation
- By conservation
- Custom highlighting

The example below highlights the sequence highlighted by standard 4 amino acid types.
Suppose a particular annotation catches our eye and we want to discover more about it. Click on the annotation of interest. A dialog comes up.

Click on "About this Annotation" and a report containing the information this particular annotation was based on will be generated and displayed in the Information Window.
**Tutorial 2: BRCA1**

*Start with manual search in Look v3*

In this example we do a DNA-based sequence study. Searching with the keyword "brca1," we discover a closely homologous family of sequences.

Go to the menu File:New:Keyword Searches and select "Sequence Databases", allowing you to search several servers using keywords. Type "brca1" and click on "Search" to launch the search. To check on progress of a search, click on the Inbox icon of the main tool bar and view the information box that comes up.

When the search is complete, matching records in the databases are loaded in the Information Window.

Click on links of interest to recover further information.
LOADING THE BRCA1 SEQUENCE

Near the top of the list you'll see the entry "gb:HSU14680 Human breast and ovarian cancer susceptibility (BRCA1) mRNA."

Hold down the mouse on the link and a context-sensitive menu will appear. Release the mouse on "Load Sequence...". Progress is indicated on the top title bar. The sequence will be read into memory and its name will appear in the "Choose Sequence(s) to Show" dialog.


**Discovery Engine User Guide**

![Choose Sequences to Show dialog](image)

**Figure 55. Choose Sequences to Show dialog**

**CHOOSE WORKUP AND PRIVACY LEVELS**

In the dialog, indicate workup level (i.e., extent of Annotations to be obtained). For this example choose "Regular." Also indicate privacy level, and whether the sequence is to be added to a new or existing alignment. You’re also given a choice of whether to align related DNA and protein sequences in the same alignment, in cases where both DNA and protein sequences are on file are involved.

When done, select the sequence and click OK. The not yet annotated sequence appears in the Sequence Window. Discovery Engine will immediately start annotating the sequence. Progress is indicated in the title bar.

![About-to-be annotated DNA sequence](image)

**Figure 56. About-to-be annotated DNA sequence**

In this example, we show annotations applicable to DNA sequences. In the figure below, you can see PCR primers and homology annotations.
**LOAD HOMOLOGOUS PROTEIN SEQUENCES**

One annotation indicates a "73% identity: breast and ovarian cancer susceptibility protein":

![DNA annotation](image)

Figure 58. DNA to protein annotation

Hold down the mouse on this annotation to bring up the annotation pop-up menu. Click on "About this Annotation." The results are displayed in the Information Window.
Hold down the mouse on the human and mouse sequences labeled "BREAST CANCER 1 SUSCEPTIBILITY GENE" and in the Information Window’s context-sensitive drop-down menu each time select "Load sequence id:...". In the next dialog, you can again select the workup level before loading these new sequences into the Sequence Window.
LOAD A TRANSLATED ORF INTO THE ALIGNMENT

One of the Open Reading Frames translated from the BRCA1 DNA sequence (+1) correlates with an extensive Homology annotation.

Figure 61. BRCA1 ORF annotation

Hold down the mouse on the annotation "ORF in Frame +1" to again bring up the Annotation pop-up. Select "About this Annotation" to show a document with both the DNA sequence and the protein sequence translated from ORF+1. The document appears in the Information Window.

Figure 62. BRCA1 ORF annotation record

Click here to load the translated amino acid sequence.
Click on the link that says "Click here to load the translated amino acid sequence." The Information Window’s context-sensitive pop-up menu appears. Click on "View sequence aa+1:16-1386".

In the next dialog that appears, choose "Workup Level: Regular" and make sure to also set "Option: Align Related DNA and Protein sequences".

Now you will obtain a sequence alignment that contains both, DNA and protein sequence from ORF in Frame 1.

Both the DNA and the protein sequence can be annotated in parallel, allowing you to cross-compare annotations obtained for the DNA and the protein sequences. If both sequences show homology to the same protein family, this fact could indicate that ORF+1 may indeed be a coding region.
2. INTRODUCTION

ABOUT THE REFERENCE MANUAL

The Discovery Engine Reference Manual is organized in eight chapters, beginning with this one:

2. Architecture & data flow: organization; data flow.
3. Calculated annotations: Content & format of calculated annotations; Scientific validation of calculated annotations.
5. Information sources: Database organization & content; Annotation sources.
7. Customizing Discovery Engine: Customizing Worldbase; Scripts; In-house PDB database; Customizing annotations output; Customizing workup levels.
8. Appendix

CONVENTIONS

Many Discovery Engine operations depend on a knowledge of functionality of the Look v3 core module. Users may find it helpful to refer to the Look v3 Reference Manual as well as this document while learning Discovery Engine operations.

Elementary operations, such as how to click and drag, how to use a word processor, etc. are not covered in this document.

The colon is used to indicate menu and sub-menu relationships, as in File:New.

Discovery Engine uses the three-button UNIX mouse. Mouse button usage may be referred to by abbreviation. LMB, MMB and RMB mean left, middle and right mouse button respectively.
A working knowledge of UNIX, while not essential, will allow you to manage Discovery Engine files and directories.
3. Architecture, Data Flow

Chapter Contents

- Organization schematics
- GeneMine architecture overview
- Discovery Engine data flow
- Sequence length constraints
ORGANIZATION SCHEMATICS

GeneMine ARCHITECTURE OVERVIEW

Figure 64. Worldbase architecture
Discovery Engine DATA FLOW

The figure gives a biological interpretation to the flow of data between GeneMine and external databases.

DATA FLOW

Queries entered at the UI undergo defined conversions in syntax and format as they proceed to the external database and then back to the UI as a query result.
Discovery Engine Reference Manual

**Queries**

GeneMine recognizes two query types: keyword and sequence. Keywords are text strings that are matched against the contents of database record fields according to instructions specified by entries in Worldbase.

**Constraints on Queries**

Non-standard characters, allowed/disallowed Booleans, wildcard symbols, case sensitivity (to distinguish DNA/RNA and Peptide), maximum sequence length, allowable IUPAC codes, gap and unknown residue characters are detected and dealt with appropriately by calls in the Worldbase file and ancillary filter scripts. Normally users need not be concerned with these matters.

**Services**

Worldbase data links are organized in "service" groups, enabling Discovery Engine to automatically recognize links to biological data and provide all possible applications to that data. For example, there is a PDB service. Whenever Discovery Engine needs to retrieve PDB information, it launches one or more SOURCES that process PDB services.

A service represents a basic category of query, e.g.

- retrieving a PDB structure by its ID code;
- performing a homology search on a sequence;
- generating automatic annotations for a sequence;
- producing a tabular report for a family of homologs obtained from BLAST.

Each service has a name, which may be used as a "filename" passed to fopen_url for obtaining results from that service without needing to know any of the details for how that service is provided. For example, in the Open URL dialog one may type `lpdb:1pk4` to load the structure 1PK4.

**Sources**

Hierarchically, sources are one level down from Services. Sources are URLs:

```
www, gopher, ftp.
```
**DATA MANAGER**

Proper security and Workup level is associated with the query at this stage of the data flow.

**SECURITY**

To protect proprietary data from being exposed to insecure network services, Discovery Engine provides four levels of security policy: public (unrestricted), regular (the site’s default policy), strict (for closely guarded proprietary data), and absolute. The system administrator for each site establishes the exact meaning of these policies via the GLOBAL-SECURITY setting in worldbase.src. Discovery Engine users then have a choice of which of the four policies to apply to any given piece of proprietary information. Discovery Engine applies the selected security policy to ensure that no information of a given security level is sent to network services with a lower security level.

Discovery Engine assigns a security value to each service, with higher values indicating more trusted services, on the following scale:

- 0-99: internet services. Less secure services are assigned a lower value.
- 100-199: intranet services, local to the site.
- 200-255: user services, i.e., scripts, programs owned by the individual user.

Each of the four security policies has a minimum security threshold setting: services below this security level will be blocked. The default threshold settings are:

- **Public**: 0. Completely unrestricted queries.
- **Regular**: 15. All but the most insecure Internet servers are allowed.
- **Strict**: 150. Only the most secure intranet and user services are allowed.
- **Absolute**: 250. Only the most secure user services are allowed.

**DESIGNATION OF SECURITY LEVEL**

By definition, all public (WWW) servers must have a security level of 0. Internal services may have security levels up to 250, as determined by the user and the sysadmin. Security level is defined in the Worldbase SECURITY= expression.
**QUEUING**

In the process of retrieving annotations, Discovery Engine’s Worldbase application may generate large numbers of hits by the processes of additive filtering, which generates multiple queries starting from a single starting URL, and event-to-service triggers, which define multiple queries to be sent out as result of single user action. These operations could jam servers and frustrate the search operation if allowed to proceed in an unregulated fashion.

Queuing and caching features resident in Worldbase prevent this from happening by:

- setting maximum number network sockets
- keeping excess queries on hold
- grouping queries into job groups
- checking queries against network cache to eliminate redundant searches (cache directory size (~/.look/cache) default size 10Mb).

Discovery Engine’s network queuing system permits automation of hundreds of simultaneous queries. The Discovery Engine infrastructure and interface provides the support for easily working with multiple simultaneous queries.

To support an unlimited number of simultaneous queries, Discovery Engine implements a simple queue system for transparently processing many queries without overloading system or network resources. It sets a maximum number of active network sockets, and keeps excess queries “on hold” until completion of other queries frees available sockets for launching them. It also goes to some lengths to track the PIDs (UNIX process identifiers) for all active network query processes so that it can control and if necessary cancel them.

In a system in which queries are processed one by one, running several queries at once could become awkward simply due to the number of alerts that might come up. Each query would be handled as an independent event and selected by the user individually to load it. There would be no way to group queries and control them as a group.

By contrast, Discovery Engine groups queries into “job” groups. Typically a job is triggered by a single user action (e.g., clicking “Load all Sequences containing this motif”) but may consist of many simultaneous queries. The job is the basic unit of control in Discovery Engine.

The In-box lists the jobs currently underway (rather than all the queries). The job is listed typically by the name of the application action the user selected to initiate.
the job. These are intended to be clearly identified and self-explanatory. Clicking a job name brings up a job control dialog, giving the current duration of the job, the number of queries completed so far for it, the total number of queries underway for it, the percent total completion of the job, and the number of errors so far. The user is given the option to let the job continue to process; stop the job and view its results so far; or cancel the job and ignore its results.

Intermediate events for individual queries within a job ("no results found", network errors, etc.) are handled silently. The user can check for errors either via the job control dialog (see above), or the Completion Notification message, which appears at the top title bar. The user can also check incoming blocks and the status of pending jobs at the title bar. In addition, all information is sent to stdout and kept in the "look_output.log" file.

Upon completion, each job has an associated final action to be performed on its results (e.g., for loading sequences, launch the Show Sequences dialog; for a document, to be shown in the browser window). If the job finishes within a short period (30 seconds) of its start, and no subsequent job was launched, the job action is immediately performed for the user. If the job took longer than 30 seconds or the user launched other jobs in the meantime, the user is notified which job com-

Figure 66. In-box pull-down and job status report

Figure 67. Real-time job report at title bar
Discovery Engine Reference Manual

pleted, and the job is placed in the In Box [IN]. Pending jobs are indicated by the In Box being highlighted in red, and by the GeneMine application beeping.

ADDITIVE FILTERING

"Additive filtering" provides flexible control for launching multiple queries from a single user action. These script-based filters can transparently spawn multiple queries from a single starting URL (e.g., from a Swiss Prot sequence ID, get the sequence entry, extract all Prosite references, retrieve them, extract all sequence references (i.e., of sequences containing these Prosite Motifs), retrieve them and load them).

EVENT-TO-SERVICE TRIGGERS

A second form of query automation launches multiple queries as a single job. For standard events in Discovery Engine (e.g., align_sequence_event: the user aligned a sequence to display it), the worldbase.src file can define a list of one or more services to be launched automatically. Thus multiple queries may be triggered automatically by user actions ("events"). In this example, all queries registered in the sequence annotation service (seq-annot-auto) would be launched as a single job titled "Annotations for name of sequence".

WWW

According to SOURCE and PATH definitions in the Worldbase.src file, a query may be routed to an internal ("intranet") or external ("Internet") source.

VIRTUAL DOCUMENTS

A query may follow links contained in an internal "virtual" document such as a keyword search result or "About this Annotation" command saved as a Hypernote. Use the Information Window’s Search... button and select "Search My Documents" in the scroll list to search virtual documents.

URL FORMATS

Network addresses may change unpredictably so they are not listed in this document. Sites are monitored continually and are updated as necessary. Current Worldbase Source addresses are available at the MAG web site:
http://www.mag.com

Approximately 100 sites and mirrors are referenced in this version of GeneMine.

**URL CHECKER**

Scripts to check the data returned by Worldbase Sources have been developed at MAG. On a high level, the information returned using sample queries to the URLs in worldbase.src is stored in a file. Then, the queries are periodically rerun to check if the URL is still returning the same data. Any differences are noted and the appropriate modifications are made to Worldbase to update correct access to Worldbase sources.

Users will be notified on a regular basis, both by e-mail and by posting to the MAG web site, of changes made to Worldbase, and updates will be made available to them.

**BACKUP SERVERS**

It doesn’t matter to Discovery Engine if the information it accesses is on the local drive or in some distant location. An alignment file could have a local filename, or an http:// address and it would be brought into the Sequence window without distinction as to its source. If a local sequence file or structure were not present in an expected location for some reason, the program’s controller (in the Worldbase.src file) as a backup will go out on the Web and look for it. The user might not even know this operation was taking place.

Discovery Engine’s backup services system copes with server failure by finding another source for performing the same query. Multiple sources (servers) may be registered for providing the same service, in either an additive or exclusive fashion. In additive access to services, all sources are queried at once and all their output is processed. In exclusive access to services, the preferred source is called first, and if it fails the next one is then called, and so on. The latter mode provides backup server capability by automatically switching to a backup server providing the same service if the primary server fails. The manner in which backup servers are called up is specified in Worldbase. The availability of dual modes for accessing backup servers makes Discovery Engine robust to the vicissitudes of the Internet.

**HISTORY STAMP**

Discovery Engine keeps a history stamp of the service it used to perform each query, as part of the recorded query path. This means the query can be regener-
CACHING

LOOK 3's caching system provides transparent query speedups, and cleans up the annoying proliferation of files that can be associated with multiple queries. Such heavy emphasis on network queries required a better infrastructure for managing and accelerating data retrieval. Like Netscape, LOOK v3 employs a "network cache" to store the results of all queries for fast retrieval in the future.

For each network query, it first checks its cache to see if it already has the results; in that case the results load immediately (and obviously no network query is performed).

The cache directory is in ~/.look/cache. LOOK v3 manages the cache directory (kept in) to keep its total size under the limit set by the user in the Network Preferences. To do so, it deletes the oldest files (old not in terms of their date of creation, but rather the date of most recent access) to reduce the cache to the required size. The default cache size is 10 Mb; a larger cache improves performance.

Also see the discussion under Edit:Preferences... in the Menus chapter later in this document.

SLOW QUERIES

By keeping all query results in the cache, Discovery Engine is not obligated to save slow query results (e.g., PDBs, BLAST searches, etc.) in the current directory. The user always has the option to save the results in the current directory; but otherwise the results are always conveniently available from the cache, with no annoying accumulation of leftover query files in the user's directory.

The caching mechanism gives the user good feedback about network performance. Discovery Engine circumvents the problem of inadequate buffer size by reading data in blocks of exactly the size of the WWW buffer, thus guaranteeing that it will never get locked up waiting in the read function. Data blocks are saved directly to cache. Real time display of total bytes read in is at the title bar message, enabling the user to gauge network activity in progress.

Users should be cautious about launching very time-consuming searches. For example, loading a large number of Entrez neighbors can take an hour or longer to
run. Ultimately there is always going to be a performance issue with "enough" sequences getting loaded, and users should be aware of this limitation.

**CACHE HASH KEY**

The hash key is based on the URL sent to the source. GeneMine uses a new algorithm to generate hash ID's. The algorithm is based on the "multiplication method" of hashing, from "Algorithms" by Cormen, Leiserson, & Rivest. This should distribute the hash id's evenly and reduce the potential for collision.

The key (i.e., the original URL) is stored in the cache file. The first line of the cache file is in the format:

```
MAG GeneMine CACHE FILE:(url)
```

where (url) is the original source of the data, with all the GeneMine/Worldbase specific information stripped. When reading in a cache file, it first reads the first line. If the "MAG GeneMine CACHE FILE" part is missing, or if (url) is different from the one it is expecting, it will not use the file. No attempt is made to handle collisions other than overwriting the old information.

**EXTERNAL FILTERING**

An external filtering layer is provided for parsing results with standard tools (e.g., Perl and Python scripting languages), which insulates internal analysis from the variability of source formatting, using completely standard parsing languages.

**APPLICATION RECOGNITION**

The worldbase.src file defines recognition filters for recognizing URLs that refer to biological data which Discovery Engine can work with (sequences, structures, etc.). These recognition filters translate those URLs to the appropriate service. A first consequence is that the service applications described above are available to any document from any source, as long as the links are recognized by the filter list.

**INTERNAL FILTERING/ANALYSIS**

Worldbase's internal filtering mechanisms allow selection of subsets of data for retrieval, and generation of new queries based on single or multiple hits from returned data, in any combination with external filtering steps.
**Sequence Length Constraints**

**By GeneMine**

Although GeneMine may be asked to perform annotations on long sequences, users should be aware that GeneMine is not intended to run with extremely long sequences. Since memory requirements for sequence alignment go as the square of sequence length, there is no way to align sequences in the range of 100,000 basepairs. Without an alignment it is probably not useful to display extremely long sequences. It is recommended that users avoid sending out queries containing more than 8,000 characters.

**By Servers**

Servers also have upper limits on the lengths of the sequences they'll accept. Following is a list of selected databases showing maximum sequence lengths that they accept. Testing in this series was performed on 17 July, 1997.

<table>
<thead>
<tr>
<th>Site</th>
<th>Documented Limit</th>
<th>Tested Limit</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alces DNA sequence translation</td>
<td>very large</td>
<td>&gt;30,000</td>
<td>webcuse not found</td>
</tr>
<tr>
<td>Prosite Scan</td>
<td>none</td>
<td>&gt;30000</td>
<td></td>
</tr>
<tr>
<td>DNA to protein</td>
<td>none</td>
<td>~2000</td>
<td></td>
</tr>
<tr>
<td>BLAST search on SCOP</td>
<td>none</td>
<td>10-15,000</td>
<td></td>
</tr>
<tr>
<td>Genquest BLASTP on PDB</td>
<td>1000</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>Genquest FASTA on SwissProt</td>
<td>1000</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>
## Discovery Engine Reference Manual

<table>
<thead>
<tr>
<th>Operation</th>
<th>Query Length</th>
<th>Subject Size</th>
<th>Time Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genquest BLASTP on SwissProt</td>
<td>1000</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>NCBI BLASTP search</td>
<td>none</td>
<td>&gt;15,000</td>
<td>takes about 25 minutes</td>
</tr>
<tr>
<td>NCBI ORF finder</td>
<td>none</td>
<td>&gt;30,000</td>
<td></td>
</tr>
<tr>
<td>NCBI BLASTN on dbEST</td>
<td>none</td>
<td>&gt;30,000</td>
<td></td>
</tr>
</tbody>
</table>
4. Worldbase

CHAPTER CONTENTS

- Introduction
- Worldbase organization
- Worldbase format
- Customizing Worldbase

INTRODUCTION

By directing outbound queries to appropriate sources, and by filtering and ordering returned output, the worldbase.src file is at the heart of both alignments and annotations. This chapter discusses the organization and variable definitions in the Worldbase file. A section at the end of the chapter provides instructions and examples of Worldbase customization.

Following chapters in this manual include numerous excerpts from the worldbase.src file to illustrate how Discovery Engine controls the dissemination of queries to accomplish particular tasks. When reviewing the file excerpts in later chapters, refer back to this chapter to as a primary information source on Worldbase.

Worldbase ORGANIZATION

worldbase.src is a text file. Worldbase entries are in blocks that begin with SOURCE=, APPLICATION= or EVENT= lines in the format of <FIELD>=<value>. 

As many as four worldbase.src configuration files may be present in different directories in a user’s copy of GeneMine. There must be at least one Worldbase, but multiple copies can be and normally are present, which typically differ between the /local (i.e., site) and /home (i.e., user) locations.

At startup GeneMine looks for worldbase.src in multiple places. As the program opens it builds a table in memory, and adds new Worldbase information every time it finds a new worldbase file.

If the SOURCE=, APPLICATION= or EVENT= variables are unique the Worldbase file gets added, although if it’s identical the most recent one overwrites the older one.

First the program looks in the LOOK_INSTALL directory. This is where the default all-encompassing MAG worldbase.src file resides; it is not user-writable. This copy of Worldbase is replaceable without compromising any existing customized files that may be present. It comes with the distribution and updates are distributed regularly by MAG as server configurations, etc., change. This contains entries accessible by everyone at the user’s site. This is the only Worldbase file that is required.

The next place the program at startup looks is in the /local directory, one level below LOOK_INSTALL. While a /local/worldbase.src file is not provided with the program, users are encouraged to establish one of these directories so that they can do their own customizations. This directory, accessible by the entire site, would be used for system wide customization. Additional worldbase files add flexibility to the use of GeneMine.

A third location potentially is in the home directory. This Worldbase might contain customizations owned by the user.

A last place that one can be is in the current working directory. A Worldbase file here is one that the user uses for testing purposes, for example seeing if a new server works; it is in effect a "scratch pad." This Worldbase would be accessible to a user from within a project.

These multiple worldbase files have the same filename but owing to the fact that they are located in different subdirectories they are distinguishable, both by the program and by the user.
Worldbase FORMAT

Databases queried by Discovery Engine store their information in records which include lines identified by Source, Pattern, Title, Keyword, etc. Each database is organized and formatted according to the standards of the agency or entity which compiled the database, and the format of one is normally not consistent with that of another. Scripts resident in Worldbase convert formats of all defined links into a common format that is recognized by GeneMine.

The Worldbase file (worldbase.src) determines how GeneMine retrieves and handles information from data sources on the internet. The file consists of single-line statements, grouped into 3 types of entries: SOURCE, APPLICATION, and EVENT. Statements are in the format of FIELD=DATA. Each of the three assignment variables can be associated with various calls, such as NAME, FILTER, PATH, etc. These calls provide Discovery Engine with the detailed information (protocol, address, data to pass, how to process the output, etc.) it needs for connecting to remote servers for data and queries.

Worldbase SERVICES

Data links are organized in "service" groups, enabling Discovery Engine to automatically recognize links to biological data and provide all possible applications to that data. For example, there is a pdb service. Whenever Discovery Engine needs to retrieve pdb information, it launches one or more SOURCES that process pdb services.

A service represents a basic category of query, e.g.

- retrieving a PDB structure by its ID code;
- performing a homology search on a sequence;
- generating automatic annotations for a sequence;
- producing a tabular report for a family of homologs obtained from BLAST.

A service is a type of information that Discovery Engine can use. Sources are registered as providing a particular service. For example, several sources can provide PDB information. Whenever a request is generated for PDB information, all sources with SERVICE=pdb are queried.

The SERVICE variable registers the source currently being defined as providing the specified service category(s). Service categories provide an abstract layer of
"service type" above the list of specific sources, that organizes them on the basis of the type of service they provide. Currently defined services in Worldbase include:

1. pdb: given a pdb code, retrieves the associated pdb file.
2. medline-uid: given a Medline Unique ID, retrieves the associated abstract.
3. swiss-seq: given a SwissProt accession code, retrieves the associated sequence file.
4. scop: given a SCOP code, retrieves the associated structural classification document.
5. prot-annot-regular: given a sequence as input object, performs queries to automatically build annotations of features for that sequence. There are several levels of annotation service (light, regular, full), as well as separate service groups for protein and dna (prot-, dna-).

Worldbase SOURCE DEFINITIONS

SOURCE

SOURCE entries describe the server connections for queries or data retrieval on the internet and provide GeneMine methods for accessing, retrieving, and parsing that data. The types of data that can be retrieved by Discovery Engine are classified into services.

SOURCE=name begins the definition of a new Worldbase source. Since this statement signals the start of a new source definition, it must be the first statement of the definition. If name matches a Worldbase source that has already been defined, the older definition is deleted, and the new definition replaces it.

SERVICE

Each service has a name, which may be used as a "filename" passed to fopen_url for obtaining results from that service without the user’s needing to worry about the details for how that service is provided.

SERVICE=service-name[,]service-name2...
To designate a backup SERVICE — that is, one which is queried only if primary services fail — by appending the & character to the end of the service name. For example, to specify a backup pdb service, use SERVICE=pdb&.

A source can register with multiple services by separating their names with & (ampersand) or, (comma).

**HISTORY**

HISTORY=service-name

When a query to a given source is performed, it is stamped with its service-name as a HISTORY field. The history stamp is often the same as the main service registration. However, a service can be registered with multiple sources. The separate history-stamp mechanism enables backup services, and provides document identification and updating.

If the query fails, another source providing the specified service can be tried instead (backup service). Also, any time in the future, the document name for the results document will contain this history stamp, both for informational purposes (to know what the document is), and for updating (the query could be redone to check for new results).

The history stamp mechanism follows a very simple rule:

- if a HISTORY statement is present, it is used;
- otherwise, if a service statement is present, the first service-name is used as the history stamp;
- otherwise the source name is used as the history stamp.

**NAME**

NAME=Description

Specifies the name of the source currently being defined, as it will appear in Discovery Engine user interface, menus, etc.

**EXAMPLE**

NAME=GenQuest search (Netscape)

**TITLE**

TITLE=Document description
When Discovery Engine performs a query, it creates a document containing the result of the search. The TITLE variable specifies the title of that document. Often, "%key%" is used to tie the document name to the keyword used in the search.

**EXAMPLE:**

```
TITLE=Entrez Search on %key%
```

**UI**

```
UI=interface-type
```

Specifies where the source will be listed in Discovery Engine’s user interface. The possible locations are

1. Keyword Search dialog: this is the default. No UI needed.
2. Sequence Queries dialog. This is the default for all sources of input type Seq. No UI needed.
3. Print Services: to register a print-source so that it appears in the menu in the printing dialog, give it an interface-type of print. UI=print.
4. Homology modeling services: to register a source so it appears in the Homology Modeling dialog, give it an interface-type of SEGMOD. UI=segmod.
5. Mutant modeling services: to register a source so it appears in the Homology Modeling dialog, give it an interface-type of CARA. UI=cara.
6. Hidden: to prevent a source from appearing in Discovery Engine’s GUI, give it an interface-type of hide. UI=hide.

**PATH**

Specifies the path pointing to the information source on the Internet. Paths can point to a URL, a UNIX program, or a Worldbase service.

Paths can utilize gopher, ftp, or http protocols. Generally, these point to CGI scripts that dynamically retrieve information.

Examples:

```
http://host/cgi-path?data - query to a CGI script using the GET method
http://host/cgi-path??data - query to a CGI script using the POST method
@http://host/cgi-path??data - returns the HTML source
```
**Discovery Engine Reference Manual**

@@http://host/cgi-path??data - interprets HTML source, strips hypertext and anchor information
@@http://host/cgi-path??data - retrieves HTML source, strips hypertext and anchor information

Paths to UNIX programs are in the format:
`| (filename) [parameters]...

Paths to Worldbase services are in the format:
`|| (service):[parameters]...

**INPUT**

INPUT=data-type

Specifies the type of input expected by the source (e.g., Seq). Normally this field is blank, except for the sources that expect either DNA or protein sequences, in which case the field is INPUT=Seq or INPUT=MultiSeq. When MultiSeq is specified, GeneMine will send the server all the sequences that the users specifies as a single query. All sources with these specifications are displayed in the Sequence Queries dialog, unless UI=hide.

**OUTPUT**

OUTPUT=document-type

Ordinarily documents are loaded and displayed within Discovery Engine’s Information Window. There are a seven other ways to load a query-result document to GeneMine. The OUTPUT variable describes how Discovery Engine should handle the data returned by the servers.

1. Molecule: to load the resulting document as a PDB-formatted structure, and display it, give it a document-type of Mol.
2. Sequence: to load the resulting document as a standard-formatted sequence file, and display it (Show Sequence dialog), give it a document-type of Seq.
3. Sequence Annotation: to load the resulting document as a formatted sequence annotation file (currently supports PIR and SwissProt formats), and display the annotations on the input sequence object, give it a document-type of SeqDBAnnotation.
4. Homology Search results: to load the resulting document as a homology search result, and display it as annotations on the input sequence object, give it a document-type of BLASTHomology.
**MAXSEQ**

Some servers that require input sequences have limits on the length of the sequence. The Maxseq value is the maximum length of the sequence that it can accept. Leaving this field blank signifies that the server can accept sequences of any length.

**AND, OR, NOT**

```
AND=and-operator
OR=or-operator
NOT=not-operator
```

Discovery Engine accepts obvious strings for logical searches. It accepts: (AND, and, &), (OR, or, |), and (NOT, not, !). But since sources on the Internet may expect a variety of different strings for these operators, these fields specify the character or string that the server expects for AND, OR, and NOT queries. Discovery Engine will replace the strings it expects with the values in these fields. Note, however, that Discovery Engine does not accept non-obvious search string operators such as "||" or "->" replacing conventional operators. In such cases it is necessary to specify the non-standard operator in the AND=, OR= or NOT= field.

Define the logical operator symbols to be used in communicating queries to this source. Discovery Engine replaces all standard logical operator forms (e.g., for logical-and: "AND" "and" "&") to the symbols defined here, so users don’t have to remember the idiosyncrasies of different servers.

**WHITESPACE**

```
WHITESPACE=character-string
```

Worldbase replaces all consecutive instances of whitespace characters (space, tab, etc.) with a single instance of the specified whitespace character; on many FORMS interfaces, the required whitespace character is +. Consecutive whitespaces are squashed into one character.

**WILDCARD**

```
WILDCARD=character
```

Discovery Engine replaces all instances of the wildcard (*) with the character specified here. Default is #.
**Discovery Engine Reference Manual**

**STRIP**

```
STRIP=character-string
```

Discovery Engine will remove all instances of the designated characters, before transmitting the query. This provides a mechanism for "working around" idiosyncrasies of individual servers, e.g., a server that dies if it sees the letter X in a sequence.

**Example**

Remove non-standard amino acid symbols X (any aa), ? (non-standard residue), ] (gap):

```
STRIP=X?]
```

A second usage of STRIP is to force query text to be in uppercase or lowercase, a requirement for some servers. If the first two letters of the character-string are AZ, uppercase letters are converted automatically to lowercase. Alternatively, if the first two letters are az, all lowercase letters are converted to uppercase.

**Example**

Convert all uppercase letters to lowercase:

```
STRIP=AZ
```

**VIEWER**

```
VIEWER=! VIEWER=document!
```

Normally, all network-based queries (and system command-based queries) are cached, so that subsequent repetitions of these queries can simply use the cached results rather than repeating potentially lengthy queries. The no-cache (!) option turns caching OFF for the source currently being defined: it will not be cached, and all queries to this source will be performed afresh.

The field is provided for compatibility purposes to older versions of Look v3. The original usage is no longer in use, and now, this field is only used to control caching. That is, VIEWER=! and VIEWER=document! are now interpreted the same.

This field is, for the most part, obsolete, but can be used to turn off caching by setting the value to VIEWER=document!
**Discovery Engine Reference Manual**

**FILTER**

FILTER=filter-command[filter-command2][filter-command3]...

When Discovery Engine queries data from an Internet data source, the format of the result is defined by the server. This field specifies the name of a script that converts the server's format to Discovery Engine's. Worldbase filters can translate any document from any source, as long as its link is recognized by the filter list.

The FILTER option provides flexibility for reformatting or reprocessing selected items from the output returned by a source. Multiple filters can be used, and will operate as a pipeline passing results from one filter to the next. When filtering is completed, the final results will be read as a document (or as whatever OUTPUT type specified for the source, see above).

Two filter types are available:

- internal filters performing simple selection or reformatting operations,
- external filters based on standard parsing languages or custom programs.

Each filter command follows an extremely simple syntax: a single letter-code specifying the type of the filter; a single delimiter character (typically the quote-mark ") for enclosing arguments to the filter; the argument string; a single delimiter character (matching the initial delimiter character), marking the end of the argument string. Discovery Engine currently supports several different types of filters:

- contains (~): searches the document for a URL which contains a match to the text argument, and opens that URL.
- contains (+): searches the document for all URLs which contain a match to the text argument, and opens all those URLs.
- Whatever further filtering steps that have been defined, and the final loading function, will be applied separately to each of the opened URLs.
- Input format (S): defines a scanf format string for reading fields from the input URL for the contains operator.
- Output format (P): defines a printf format string for printing fields to the output URL for the contains operator.
- awk (a): runs the text argument as an awk script, with the current document as input on stdin. If the text argument begins with $Discovery Engine_DIR,
it treats it as a path to an awk script stored in the Discovery Engine installation tree.

- perl (p): runs the text argument as an argument to perl, with the current document as input on stdin.
- system command (s): runs the text argument as a system command, with the current document as input on stdin.

**EXAMPLE 1**

Select the first URL in the document containing the phrase "db/release/ddbj", and load that document:

```plaintext
FILTER=~"db/release/ddbj"
```

**EXAMPLE 2**

Parse the document using an awk script stored in the Discovery Engine install:

```plaintext
FILTER=a"$Discovery Engine_DIR/exec/eerie-fasta.awk"
```

**Worldbase APPLICATION DEFINITIONS**

An application entry defines an action (e.g., "Load this sequence") to be offered as a menu item for URLs matching the specified pattern. The GeneMine worldbase.src file can define any list of applications which can be performed on data belonging to a given service. Applications represent URLs which provide data that Discovery Engine can use in a more powerful way than Netscape can (such as PDBs, sequences, etc.). Discovery Engine highlights these links with a hot hyperlink color (red), and shows a pop-up menu of applications available if the user clicks on such a link. Discovery Engine recognizes URLs for which it has applications by simple text rules. Discovery Engine’s applications are generally defined in two layers: first, to recognize URLs that link to a known kind of data (e.g., gopher://pdb.pdb.bnl.gov/00/PDB/Entries/.4pti/4pti.full, a PDB file), and reformat that URL into a service reference (e.g., ||pdb:4pti); and second, applications available for each kind of service (e.g., for the pdb service, "View this Structure"). By separating the recognition of data types from the definition of applications for those data types, we avoid duplicating our application definitions endlessly. The first layer specifies only a pattern to recognize, and a new path (generally to a Worldbase service) to apply in place of the original path. Applications in the second layer specify in addition a label for listing this application on the pop-up menu.
GeneMine applications can be customized to provide whatever action is desired (e.g., retrieve and display a structure as one action) with an appropriate explanatory label.

Any number of applications can be provided, and are displayed on the applications popup menu. For example, for hyperlinks referencing structure, GeneMine provides the following applications: View (retrieve and display); Get Info; Look v3 up in LinkDB; Get Structural Classification (from SCOP), etc.

The following fields define APPLICATION entries. The entries are in the format: APPLICATION=(source name). The fields can be specified in any order, one field per line.

**APPLICATION**

APPLICATION=

This is the name of the application.

**PATTERN**

PATTERN==

PATTERN==

PATTERN=%

Describes the URL addresses that this application recognizes. The format of this is:

(type code)(recognition string)

The type code can be one of the following values:

== Exact Match. The beginning of the URL must match the recognition string.

= Contains. The URL must contain the recognition string.

%: sscanf Match. The recognition string must be: (number of arguments):(sscanf format string). The argument count indicates how many fields must be matched by sscanf, reading from the URL as input; a colon; the rest of the pattern is treated as a sscanf-format string for reading the URL. The URL matches the pattern only if

(count <= sscanf(URL, format, arg1, arg2, arg3, arg4, arg5...))

This provides a flexible way of matching URLs, requiring any number of variable arguments of a given form. To use the sscanf call, one should be very famil-
iar with scanf formats: features commonly used here include %s, %d, %f, %[^exclude-set], %s, %23s, %2$[a-z_] etc. If any of these are not thoroughly familiar to you, see the scanf man pages, as these and many other scanf tricks are essential tools here. Up to 8 separate arguments may be used.

Scanf is made to do double-duty: first, to test whether the URL matches a desired pattern; and second, to read desired fields from the URL into an argument list, which later can be used to generate a new path or label for the application (see below).

**PATH**

```Path=
```

This specifies a path for the new query of the recognized pattern. This is a printf format string that uses the arguments recognized by the pattern.

**LABEL**

```Label=
```

When the user click-and-holds a red link, a menu pops up with several options. LABEL specifies, in a printf format string, how the application will appear in that menu.

**OUTPUT**

```Output=
```

As for SOURCE.

**FILTER**

```Filter=
```

As for SOURCE.

**APPLICATION ENTRY EXAMPLES**

The label-format is a printf format-string used to generate the label exactly as for the path (see above).

**EXAMPLE 1: VIEW ANY PDB STRUCTURE:**
**Discovery Engine Reference Manual**

APPLICATION=pdb-view  
PATTERN=%1:||pdb:%s  
OUTPUT=Mol  
LABEL=View Structure %s  

**EXAMPLE 2: LOAD ALL SEQUENCES REFERENCED TO AN EC NUMBER**

APPLICATION=ecenzyme-seqs-get  
PATTERN=%1:http://expasy.hcuge.ch/%*[\^b]bin/get-enzyme-entry?%s  
FILTER=+/"sprot-entry"  
OUTPUT=Seq  
LABEL=Load Enzyme Family %s  

**Worldbase EVENT DEFINITIONS**

For standard events in Discovery Engine (e.g., align_sequence_event: the user aligned a sequence to display it), the worldbase.src file can define a list of one or more services to be launched automatically. Thus multiple queries may be triggered automatically by user actions ("events"). In this example, all queries registered in the sequence annotation service (seq-annot-auto) would be launched as a single job titled "Annotations for name of sequence".

EVENT=show_sequence_event  
EVENT=search_worldbase_event  
EVENT=start_look_event  
EVENT=changed_align_list_event  

The Worldbase configuration in the current release of Discovery Engine relies on EVENT= definitions only for internal operations, and users normally will not have occasion to change them.
CUSTOMIZING Worldbase

The primary function of the worldbase.src file is to provide run-time definition of the servers and services used by Discovery Engine. This allows system administrators to localize the program to work with their site’s preferred resources, and extend Discovery Engine at any time with functionality from new, external services. This is done by defining a SOURCE for each type of information connection to be used by Discovery Engine. This definition specifies:

1. the protocol and path for connecting to the server.
2. the data to send to the server, and their precise format.
3. how to represent this source within Discovery Engine’s user interface.
4. how to read and present the output returned by the server.
5. each source may be defined as belonging to one or more service class, which represents its generalized query / data category (such as "PDB structure data").

DEFINE NEW SOURCE

Example: begin definition of a new source, to be called "Genquest-NS".

SOURCE=Genquest-NS

SPECIFY PATH TO SOURCE

PATH=URL

Specifies the full path to the source, including its protocol, address and the data to send it. Worldbase paths are a superset of the Uniform Resource Locator syntax, including capabilities to access system commands, internal Discovery Engine functions, or a Worldbase service (see below). The syntax of the different paths is:

STANDARD RETRIEVAL

  gopher://host/directory.../file
  ftp://host/directory.../file
  http://host/directory.../file
**KEYS**

The Worldbase format provides a variety of ways to designate data to be passed in a query.

%key%

The text keyword(s) passed to the query, either from the Keyword Search dialog, or from a Worldbase service request.

**EXAMPLE: KEYWORD SEARCH OF NIH GOPHER SERVER**

PATH=gopher://gopher.nih.gov/77/gopherlib/indices/pdb/index%key%$Discovery Engine_DIR

**SOOP FIELDS**

The proprietary Soop scripting language can be used to insert sequence information into a query.

printf-formatted SOOP fields: <<<format-string>>>

The expressions inside the <<<...>>> characters specify the operations or definitions to be included in the query.

<<<%name%s>>> - passes the sequence name to the query

<<<%seq%s>>> - passes the sequence to the query

<<<%title%s>>> - passes the title to the query

<<<%comments%s>>> - passes user comments (editable in Edit Sequence Info) to the query

<<<%accession%s>>> - passes the record accession number to the query

<<<%organism%s>>> - passes the name of the organism to the query

<<<%journal%s>>> - passes the name of the journal to the query

<<<%article_title%s>>> - passes the article title to the query

<<<%authors%s>>> - passes the author name to the query

**GET QUERY**

Performs a GET query using the specified http server.
http://host/cgi-path?data

**POST QUERY**

Performs a POST query using the specified http server.

http://host/cgi-path??data

Retrieves the specified document / query result in the original source representation (typically HTML, but this varies according to the server and specific document), rather than interpreting the HTML. This is useful primarily for data read directly by Discovery Engine internal functions.

@http://host...

Retrieves the specified document / query result in a flat text representation, interpreting the HTML but stripping all hypertext and anchor information from the output. This is useful primarily for data read directly by Discovery Engine internal functions.

@@@http://host...

Retrieves the specified document / query result as the HTML source and strips all the hypertext and anchor information. This differs slightly from @@ because here, the HTML is not interpreted.

**SYSTEM COMMAND**

| command [parameters]...

Runs the UNIX program command with the given command-line parameters, and reads its stdout output.

**SEARCH BY SERVICE**

||service-name:[parameters]...

Searches the Worldbase source definitions table for a source or sources that provide the named service. If found, that source is used, with the specified parameter list passed as the text argument to the source (see the Worldbase data format description, below). If no source matching the service is found, no query is performed. In doing the search, all sources are checked for an exact match against their service list (see below); if none match, they are checked for a match against their source name.
Discovery Engine Reference Manual

EXAMPLE: RETRIEVE ITEM 4PTI FROM THE PDB SERVICE:

| pdb:4pti |
5. Calculated Annotations

Chapter Contents

- Introduction
- Calculation method
- Format rules
- Annotation display
- The 6 calculated annotations
- Validation of method

Introduction

Six of the annotation categories reported by Discovery Engine are produced by considerable client-side processing to produce the final annotations from the original BLAST/FASTA output from external servers. The six categories are:

- Sequence Homology
- Sequence Polymorphism
- Structural Homology
- Close Homology
- EST Expression
- Homology Fingerprint

In this chapter we discuss the calculation methods used to produce the annotations, format rules and display features. We then discuss each of the calculated annotations in detail, and conclude with an overview of a study conducted at MAG to confirm the accuracy of calculated annotations.
CALCULATION METHOD

When given a new query sequence, Discovery Engine searches for homologs using a variety of sources, including SCOP BLAST, SwissProt FASTA, and dbEST. The outputs from those sources are standardized using filters, so that the results of these services can be interpreted in the same way. Discovery Engine clusters the significant homologs into families and calculates annotations based on these families.

HOMOLOG SIGNIFICANCE

Sequence search engines return a p-value for each hit. This statistic represents the probability that the match between the query and database sequence could have been derived by chance. A common and useful way of representing this probability is by a log odds (lod) score. The lod score, the negative logarithm of the p-value, has several advantages because it is easy to represent, interpret, and calculate.

\[
\text{lod} = -\log_{10}(p\text{-value})
\]

As a cut-off for significance, Discovery Engine only considers homologs with lod values that are \( \geq 0.3 \) (p-value of 0.50). For each cluster, it tracks a weighted sum of the lod scores for each sequence. It ignores the clusters when the weighted sum of the lod scores is less than 1.5.

The default family cutoff value of 0.50 is user-adjustable in the Edit:Preferences...:Annotations... dialog.

RAW HOMOLOGY SCORE RESULTS

Raw results of homology searches can be visualized by clicking on the link in the Information Window. Shown below are summary, histogram, and individual alignment results of a FASTA search.
Figure 68. Raw homology score summary

Figure 69. Raw homology score histogram
Discovery Engine Reference Manual

**BLAST ALIGNMENT SCORING**

BLAST1 doesn't handle gaps, so it often outputs multiple hits per sequence, representing the different pieces of an overall alignment with gaps between them. The BLAST server returns the results in order of highest scoring pieces and Discovery Engine then reassembles the pieces. For each alignment, BLAST supplies one character per residue describing the quality of the match between the aligned amino acids. When two sequence fragments overlap, Discovery Engine uses the match from the most recently read alignment. This scoring anomaly is in the nature of the BLAST algorithm; Discovery Engine correctly lists the results in the order the BLAST analysis ranks them.

Note that this aspect of BLAST alignment scoring is relevant to BLAST 1 only, not to any FASTA or Smith Waterman server. GeneMine also accesses the BLAST 2 server at NCBI. BLAST 2 does handle gaps and therefore the results of BLAST2 searches do not have the confusing complexity of BLAST1 searches where overlaps are present in the alignments. To perform a BLAST2 search go to Sequence:Queries and choose "NR BLAST2 Homology Annotations."

**OUTPUT PROCESSING**

Because raw BLAST output files tend to be verbose, cryptic, noisy and redundant, GeneMine applies automated routines to process output to make it more comprehensible. First a perl script parses out the variations by the different servers to present the output in a standardized format.
Next, filtered text is parsed into hits by GeneMine. Homologs are screened for significance by score ranking. Ideally the scoring is based on a FASTA p-value, when available. GeneMine tries to work with log-odds type scoring.

Next, homologs are clustered into families by one of two criteria. For structural data, they are clustered according to the Structural Classification of Proteins (SCOP) system into protein superfamilies. Homologs in the same protein superfamily are treated as a single cluster. Sequences which have no linkage to structure are clustered by sequence similarity, into families of >50% sequence identity within each family. Because the BLAST/FASTA results return alignments of source sequences to the query sequence, these alignments can be used to directly compute the similarity of the homologs without having to perform Needleman Wunsch alignment calculations on all possible pairs.

This clustering reduces the volume of distinct hits many-fold, typically three-to ten-fold. GeneMine displays each family as a single line annotation, rather than multiple lines for every homolog.

GeneMine uses log-odds scores for assessing the homologs, and thus can sum the log-odds scores for multiple homologs found to form a single family to identify families of high significance. This is especially significant when performed over SCOP superfamilies.

GeneMine also characterizes families by performing lookups to obtain biologically meaningful information for the family, e.g., "Urokinase-type protease" instead of the uninformative "P09784: PSTR_CANFA CANUS FAMILIARIS".

**HOMOLOGY HIT FILTERED OUTPUT**

An example of Discovery Engine's homology hit output is shown below. An NR FASTA Homologies (IDEAS) search (full workup, regular security level) was run on the 1hlbA sequence from 1hlb.pdb. The results were passed through Discovery Engine's ideas-fasta.pl Perl filter. The search returned over 800 hits. The beginning and end of the output is shown.

```
SIMILAR-CHARS=.
INFOLINK=||homology-note: "%1$s %2$s %4$s %8$s %6$s\nt=PIR Homology Family
%1$s

SEQ-ID=GLBC_CAUAR
```
Discovery Engine Reference Manual

LABEL=GLOBIN C, COELOMIC.
DATABASE=swiss-id
EXPECT=0
QUERY-POS=1
QUERY=GGTLAIQAQGDLTLAQKKIVRKTWHQLMRNKTSFVTDVFIRAYDSPAQNKFPQAGMS
HOMOL=:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SUBJECT-POS=1
SBJCT=GGTLAIQAQGDLTLAQKKIVRKTWHQLMRNKTSFVTDVFIRAYDSPAQNKFPQAGMS
QUERY-POS=61
QUERY=ASQLRSSRQMQAHAIRVSSIMSEYVEELSDILPELLATLARTHDLNKVGADHYNLFAKV
HOMOL=:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SUBJECT-POS=61
SBJCT=ASQLRSSRQMQAHAIRVSSIMSEYVEELSDILPELLATLARTHDLNKVGADHYNLFAKV
QUERY-POS=121
QUERY=LMEALQAELGSDFNEKTRDAWAKAFSVVQAVLLVKG
HOMOL=:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SUBJECT-POS=121
SBJCT=LMEALQAELGSDFNEKTRDAWAKAFSVVQAVLLVKG
END-HIT

SEQ-ID=A53881
LABEL=hemoglobin chain C - sea cucumber
DATABASE=genbank-prot
EXPECT=0
QUERY-POS=2
QUERY=GGTLAIQAQGDLTLAQKKIVRKTWHQLMRNKTSFVTDVFIRAYDSPAQNKFPQAGMS
HOMOL=:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
SUBJECT-POS=2
SBJCT=GGTLAIQAQGDLTLAQKKIVRKTWHQLMRNKTSFVTDVFIRAYDSPAQNKFPQAGMS
QUERY-POS=61
QUERY=ASQLRSSRQMQAHAIRVSSIMSEYVEELSDILPELLATLARTHDLNKVGADHYNLFAKV
Discovery Engine Reference Manual

SEQ-ID=JT0624
LABEL=hemoglobin alpha 1 chain - Komodo dragon
DATABASE=genbank-prot
EXPECT=9.4
QUERY-POS=11
QUERY=DLTLAQKKIVRKTWHQLMRNKTSFVTDVFIRIFAYDPSAQNKFPMAGMS
SUBJECT-POS=1
SBJCT=VLTEDDKTHVKTLWGHVHNAEEIAADALTRMPLAHPTSKYF----AHFD
QUERY-POS=61
QUERY=ASQLRSSRQMQAHIRVSSIMSEYVEELSDILPELLALTARHTDKNGVDHYNLFAKV
SUBJECT-POS=48
SBJCT=FSP--NSANIKAHGKKVANALNQAVNHLD-----DIGGTLSDKLSDLHAQQLRDPVNGF
END-HIT

SEQ-ID=AF020281
LABEL=CbpB [Dictyostelium discoideum]
DATABASE=gp-acc
HOMOLOG ALIGNMENT

Each of the homology sources provide sequence alignments between the query sequence and found homologs. Using this information, Discovery Engine classifies each residue position of the aligned query and homologous sequence as either identical, similar, mismatched, or deleted. Then, it assigns a score to each position, based on that classification. This score is used for some of the annotations.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical</td>
<td>3</td>
</tr>
<tr>
<td>Similar</td>
<td>1</td>
</tr>
<tr>
<td>Mismatched</td>
<td>-1</td>
</tr>
<tr>
<td>Deleted</td>
<td>-3</td>
</tr>
</tbody>
</table>

Two identity scores are listed: the first is the percent identity of the query to the hit, and the second is the percent identity of the hit to the cluster leader.

The significance of the family is the sum of the lod scores for each homolog in a family.

The homology of a query sequence to a cluster is the highest sequence identity of the query to any sequence in the cluster.
The maximum number of families allowed is 100.

**HOMOLOG CLUSTERING**

In order to reduce the amount of information returned by homology searches, Discovery Engine groups the homologs into families. This clustering is done based on the structural information returned by SCOP. Discovery Engine creates families based on the SCOP superfamily classifications of the homologs.

If no SCOP information is available, then Discovery Engine adds a new homolog into a family with which it has a sequence identity greater than a clustering threshold, which is set at 50% by default. The sequence identity is the percentage of identical amino acids in the alignment between two sequences. In Discovery Engine, this number is only calculated if the number of aligned amino acids is at least 20. Otherwise, the percent identity is 0.

The highest-scoring homolog to the query that is returned by the BLAST server is made the "leader sequence." Other sequences homologous to the leader are grouped with it. That is, the leader is similar to the query, and the other sequences are similar to the leader.
**HOMOLOG OUTPUT**

GeneMine reports hits in the order in which the search server (BLAST or FASTA) ranks them. A typical output example is shown below:

Homology Family: orphan nuclear receptor - Rattus nor

This family represents a cluster of closely related sequences which were found to be homologous to your sequence 7UP1_DROME.

Click here for the Search Results Report for 7UP1_DROME

<table>
<thead>
<tr>
<th>SCORE</th>
<th>NAME</th>
<th>%IDENTITIES</th>
<th>#RES</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.8</td>
<td>2305205A</td>
<td>28</td>
<td>100</td>
<td>orphan nuclear receptor - Rattus nor</td>
</tr>
<tr>
<td>5.6</td>
<td>MUSSHHP</td>
<td>27</td>
<td>87</td>
<td>shp gene product [Mus musculus]</td>
</tr>
<tr>
<td>5.3</td>
<td>HUMSHP</td>
<td>29</td>
<td>83</td>
<td>nuclear hormone receptor [Homo sapiens]</td>
</tr>
</tbody>
</table>

Click and hold the left mouse button here to `LINK:||load-seq-set: 'prf-acc:2305205A' 'gp-acc:MUSSHHP' 'gp-acc:HUMSHP'` select sequences to display from this complete set.

Notes:

1. The score is an expectation value expressed as a log-odds score. Thus a score of 1 means a 10**-1 probability (10%) that the quality of match could have occurred by random chance. Similarly, a score of 2 means a 10**-2 probability (1%). Thus, scores below 1 may not be significant hits.

2. Identity scores are given for each sequence's homology to the *query* (7UP1_DROME), and then to the *top-scoring sequence* of this cluster. The top-scoring sequence will therefore always have a cluster value of "100".
3. #RES is the number of residues from each sequence which were aligned to 7UP1_DROME by the search to calculate the score.

4. Occasionally cluster members may appear in unusual log-odds score order due to anomalies in BLAST 1.x overlapping hits. This does not occur with BLAST2 or FASTA searches.

The reason the scores are not sorted in descending order is that the search server arrives at its final scores through the intermediate steps of init1 and init2 scores. These intermediate scores are the basis for ranking of sequences that have uniformly high final scores. To avoid floating point blowup, GeneMine truncates the zero score (which would give a log-odds score of -infinite) to a very small number 10^-31.

View scoring details by clicking on the blue "Search Results Report" hyperlink in the Information Window.

**FORMAT RULES**

**ANNOTATION FIELDS**

GeneMine can process 2 different types of annotations: BLASTHomology and SeqDBAnnotation. When a request for an annotation is sent to a server, the results are returned and then processed by a script into the format expected for the annotation type.

**ANNOTATION TYPES**

BLASTHomology searches a sequence database and returns the sequences homologous to the query sequence.

SeqDBAnnotation contains information about general types (including Prosite, Prodom and SecStr) of annotations on the sequence. SeqDB-type annotations are discussed in the following chapter, "Standard Annotations."
BLASTHOMOLOGY

The basic formatting rules are fairly simple:

1. a formatting header, giving the formats to be used for reading the homology results.
2. a series of homology hits, each giving a name, score, and an alignment.

Query sequences are sent out to servers that run searches against databases and then return alignments to homologous sequences. GeneMine reads these alignments, performs some calculations, and then displays their families in the sequence window. In addition, it has the ability to create a hypertext that displays information about the family when a user clicks on it and selects "About this annotation." In the worldbase.src, these types of servers are called BLASTHomology.

BLASTHomology servers return a series of alignments with the identical residues, similar residues, mismatched residues, and gaps labelled. For each alignment, the statistical significance of the match is typically given as a p-value or expect value.

BLASTHomology results are separated into two sections. The header gives general information about how to interpret the hits. The Hits section is a list of all the hits that are detected.

**BLASTHOMOLOGY HEADER FORMATS**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MINIMUM-SCORE=score</td>
<td>This is the minimum log-odds score of a homolog that GeneMine will consider as significant. GeneMine will not display alignments with scores below this value. By default, this value is 0.3.</td>
</tr>
<tr>
<td>FAMILY-SCORE=score</td>
<td>This is the minimum log-odds family score that GeneMine will consider as significant for a family. The family score is the sum of all the log-odds significances of the members of the family. By default, this value is 1.5.</td>
</tr>
</tbody>
</table>
MISMATCH-CHARS=chars These are the characters in the alignment that specify a mismatched pair of residues. By default, GeneMine expects ' ' to signify mismatched residues.

SIMILAR-CHARS=chars These are the characters in the alignment that specify a similar pair of residues. By default, GeneMine expects '+' to signify similar residues.

GAP-CHARS=chars These are the characters in the alignment that specify a gap. By default, GeneMine expects either '-' or ' ' to signify gaps.

FAMILY-IDENTITY=percent-identity This is the minimum sequence identity required to add a new sequence to a family. See the worldbase annotations documentation for more information. By default, this value is set at 50%.

DATABASE=database-name This is the name of the database that is the source of the homologous sequence. This field can also be given before each hit for sequences that are from different databases. When the user requests more information about a cluster, i.e., clicking on "About this Annotation," GeneMine generates a hypernote that lists all the sequences in the family. Each sequence has a hyperlink to the database, with the accession number as a parameter. For example, if a pdb protein 1hlb is in a family, then that family's hypernote will contain a link to "|pdb:1hlb." In this case, pdb is the database name.

BLASTX-MODE Insert this line to signify that this report was generated by a DNA search against a protein database. GeneMine will translate the amino acid sequence positions given in the output into nucleotide sequence positions.
**Discovery Engine Reference Manual**

**EXPRESSION-MODE**
Insert this line to signify that this search, either DNA or protein, was done against an EST database. GeneMine will look for expression pattern words to add onto the label.

**SEQREF-FORMAT=FORMAT-STRING**
Search results are displayed in a hypernote. This is a printf string that specifies how the results should be formatted. If left blank, GeneMine uses "%%l%%l%%l%.1f%%l%.3f%%l%.3f%%l%.3f" as default. This field should not be changed.
INFOLINK=path

This provides information for linking a hypernote to the hits found in this search. Several fields may be used:

%1$s: the title of the homologous sequence, as obtained from the family service.
%2$s: the name of the homologous sequence, as supplied by the homology search result.
%3$s: the title of the homologous sequence, as supplied by the homology search result.
%4$s: the name of the query sequence.
%5$s: the title of the query sequence.
%6$s: the list of sequences in the homology family, supplied with score statistics, quoted (i.e., enclosed in quotes to allow internal white space) and separated by white space. For each sequence, a set of eight values is printed separated by colons:
the name of the homologous sequence, as supplied by the homology search result.
the title of the homologous sequence, as supplied by the homology search result.
the score of its alignment versus the query sequence.
its percent identity versus the query sequence.
its percent identity to the master (top) sequence of the homology family.
the length of its alignment with the query sequence (# of residues).
For DNA sequences, the amino acid translation.
The name of the report-service to apply for annotations generated from this homology.

Table 2. BLASTHomology header format
**BLASTHomology Hits Format**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEQ-ID=id</td>
<td>This is the id (typically an accession number) used to retrieve this sequence from a database.</td>
</tr>
<tr>
<td>SCOP-ID=scop-id</td>
<td>This field is given if the sequence is classified within the scop database.</td>
</tr>
<tr>
<td>LABEL=label</td>
<td>This is the label placed on the annotation for the homolog hit.</td>
</tr>
<tr>
<td>EXPECT=p-value</td>
<td>This is a p-value for the significance of the hit.</td>
</tr>
<tr>
<td>SCORE=score</td>
<td>This is a log odds score of the significance of the hit. Either expect or score must be provided.</td>
</tr>
<tr>
<td>IDENTITY=percent-identity</td>
<td>This is the percent identity of the hit. If this field is not given, GeneMine will calculate its value. Optional</td>
</tr>
<tr>
<td>READING-FRAME=frame</td>
<td>DNA Specific This is the reading frame (1, 2, 3, -1, -2, -3) returned by BLASTX searches. When searching a DNA sequence against a protein database, the query sequence must be translated into an amino acid sequence. GMP needs to know the direction the putative ORF runs along the query DNA sequence and the frame of the codon-start position.</td>
</tr>
<tr>
<td>STRAND-DIR=dir</td>
<td>DNA Specific This is the direction (1, -1) of the query sequence that matches the homologous sequence. This is given by BLASTN search results.</td>
</tr>
<tr>
<td>QUERY-POS=sequence-position</td>
<td>This is the residue number of the first aligned amino acid of the query sequence. QUERY-POS must go before the QUERY= line.</td>
</tr>
</tbody>
</table>
**Table 3. BLASTHomology hits format**

**QUERY=sequence**
This line gives the fragment of the query sequence for which the homology was found.

**HOMOL=sequence**
This is a string that specifies the alignment between the fragment of the query sequence and the fragment of the homologous sequence. Each character in this string specifies whether the residues are identical, similar, dissimilar, or gap.

**SUBJECT-POS=sequence-position**
This is the residue number of the first aligned amino acid of the homologous sequence. SUBJECT-POS must go before the SBJCT= line.

**SBJCT=sequence**
This line gives the fragment of the homologous sequence. This should be the same length as the Homol and Query sequences.

**END-HIT**
Since homology servers split single hits into several lines, GeneMine will accept multiple QUERY, HOMOL, SBJCT fields per hit. END-HIT indicates that no more fragments will be given for this hit.

**EXAMPLE HEADER**
For a set of Swiss-Prot sequences, obtain a sequence name from the swiss-seq service using the Description (DE) line, and link each homology annotation to the swiss-note service, passing it the name of our family, its ID, the name of our query sequence, and the list of homologs in the family, with statistics.

SIMILAR-CHARS=.
MINIMUM-SCORE=80
FAMILY-SCORE=100
INFOLINK=|swiss-note:"%1$s" %2$s "%4$s" %6$s#t=Swiss-Prot Homology
Family %1$s
Discovery Engine dynamically recompacts the annotation display to squeeze out any empty space between the annotations that are shown. This makes the displays much more manageable and viewable. To give the user total control of the avalanche of information returned by Discovery Engine, there are display preferences for what types of annotations are initially displayed (whether they are collected is a separate issue, controlled by the workup level). Again, separate control panels are given for "homology group leaders" (i.e., not homologous to anything else we’re looking at), vs. very close homologs (which would be pretty redundant information). These controls are at the individual annotation type (the current version of Discovery Engine contains 21 types), so the user can really customize the profile of info they want to see. Control of many feature settings through a Preferences dialog, including annotation display defaults, choice of alignment matrices, font sizes for sequences and hypertext, width of bonds and ribbons, etc.

GAP SPANNING

By default, Discovery Engine does not break up an annotation (i.e., display it as two separate annotation marks) if a short gap is present in the alignment. Sequence numbering is discontinuous, however, when gaps are present.

THE 6 CALCULATED ANNOTATIONS

SEQUENCE HOMOLOG

Sequence Homology displays families derived using sequence identity as a single line for each family. Families comprised of a single sequence, or multiple sequences that have very close sequence identity, by default SINGLE_GENE_IDENTITY_MIN=95%, are shown with a cyan line. Families of multiple sequences are shown with a white line.

Sequence homology annotations are obtained through these steps:

• Sequence similarity search by BLAST/FASTA.
Low-stringency filtering of individual hits for significance.

Reduce data bulk by clustering into homology families, using structural superfamily classification, when available, or the 50% identity rule.

50% identity rule

By the 50% rule, any sequence with greater than 50% homology to the scoring sequence ("lead sequence") of a family joins that family. To view the lead sequence of a homolog family, click "About this Annotation"; the lead sequence is the one listed at the top of the Information window. Sequence identity is defined as the percentage of identical amino acids in the alignment between two sequences. At least 20 amino acids must match, regardless of sequence length, for sequence identity to be calculated. Sequences with fewer amino acids will be reported as having 0% identity. The maximum number of families allowed is 100. Two identity scores are listed: the first is the percent identity of the query to the hit, and the second is the percent identity of the hit to the cluster leader.

From these clusters, both "Most Conserved" sites and "Least Conserved" sites (see also "Sequence Polymorphisms") are calculated. A Most Conserved region
Discovery Engine Reference Manual

is a highly conserved sequence region that is likely to be associated with a functional element. A Most Deleted region is a non-conserved region of sequence that is likely to be associated with a loop at the surface of the protein. See also "Homology Fingerprint" and "Sequence Polymorphisms."

Sequence identity is defined as the percentage of identical amino acids in the alignment between two sequences. At least 20 amino acids must match, regardless of sequence length, for sequence identity to be calculated. Sequences with fewer amino acids will be reported as having 0% identity. The maximum number of families allowed is 100.

HOMOLOG MAPPING

For Most Conserved homology, GeneMine maps features of the query sequence to all of the homologs judged to be significant. The top 10% (most conserved regions, probably representing regions critical for function); and the bottom 10-25% (least conserved=most deleted regions, probably representing putative loops).

In order to avoid false positives that can arise from incomplete data, GeneMine uses a sophisticated approach for mapping these homologous regions.

For each sequence position, several types of states (identical, similar, mismatched, deleted) are tracked. The lowest and highest levels of overall pair identity at which they occur in the entire set of homologs are recorded. For example, if a given residue was conserved in all homologs, even in the most distant homolog (e.g., 30% overall identity), that shows this residue is strongly conserved. This approach is conservative in that it finds positions which were always observed to be conserved even in distant homologs. GeneMine annotates the top 10% of these conserved positions as "most conserved homology" (blue + annotation).

Conversely, GeneMine tracks the highest-level of pair-homology at which a given position was observed to be deleted, and annotates the top 10% of these positions as putative loops (blue X annotation). If a position was observed to be deleted between two sequences that are closely related, it’s likely to be in a loop.

To keep fingerprint annotations with the associated homology annotations during the dynamic recompaction phase of output processing, GeneMine introduces annotation linking. An annotation can be linked to another annotation, such that it will always be displayed with that annotation.
VIRTUAL DOCUMENT LINKS

Homology annotations are linked to virtual-documents that provide a report-style summary of the search statistics for the family. "Virtual document" means it is not stored as an actual text document, but is produced on the fly when the user requests it. For the SCOP structural homolog annotations, this is provided by the scop-note service and lists a table of the homologs found within that SCOP superfamly. The note contains links to view the original SCOP search text output, to load the structures of the homologs, and to the SCOP web page for that superfamily.

Swiss Prot homology annotations are linked to a similar style report, produced by the swiss-note service. This report lists the statistics for the homologs, and provides hyperlinks to view the original FASTA text output, load the sequences, or get other detailed information about them.

INFOLINK

Click "About this annotation" to view statistical and other supporting data in the Information Window. Sequence Homolog annotations are derived from the family tables derived from BLAST/FASTA output.

![Figure 72. Infolink prompt and returned results](image-url)
**Default display style**

Homologies are indicated by a single line for each family. Families comprised of a single sequence, or multiple sequences that have very close sequence identity (by default 95%) are shown with a cyan line. Families consisting of more diverse sets of sequences are shown with a white line.

Families are labeled: identity% Identity: swissprot name. Text adjacent to line identifies homolog, percent identity, and (in parentheses) number of proteins in homologous cluster (e.g., 39% identity (8): DNA POLYMERASE pir).

The percent identity refers to the similarity of the query sequence to a particular member of a close homolog family. Thus, while the homologs are all closely similar within their family, they may not be closely similar to the query sequence. For this reason, low percentage scores may be present in homology annotations.

![Hemoglobin (Sea Cucumber) Display Style](image)

**Figure 73. Display style for sequence homolog**

**SEQUENCE POLYMORPHIC REGIONS**

Sequence positions where deletions or mutations are observed are annotated. Mutation Annotations are extracted from the protein file. Deletions are calculated from the homologs. Deletions are commonly observed at these positions, and may indicate the presence of surface loops.

At each residue position of the query sequence, a deletion score is calculated. This score is the maximum percent identity of all homologs with a deletion at that residue. Discovery Engine marks the 10-25% of the residues that have the highest deletion scores as Most Deleted.
**Discovery Engine Reference Manual**

**10-25% RULE**

These top 10-25% of the least conserved residues are marked polymorphic. Thus, if there are 50 aligned residues, it marks the top 5-12.5 (10-25% of 50) most deleted residues as polymorphic. For example, if 15 residues all get the highest polymorphism score, then none of them are marked because the number of residues doesn't fall in the 10-25% range. If 2 residues get the highest polymorphism score, then GeneMine will look at the next highest polymorphism score. Then, if 7 residues get the next highest score, it will mark those 9 (2 + 7) residues as being the least conserved.

This range cannot be changed by the user.

**Default Display style**

Cyan X character and text identifier for most deleted regions (putative loops) (e.g., *Most deleted in Genbank NR protein homologs*), italicized to indicate uncertainty.

Blue ▲ character to indicate known mutations (e.g., M -> T (IN LFS))

---

**HEMOGLOBIN (SEA CUCUMBER)**

```
QKKIVRKTWHQLMRNKTSFVTDVFIRIDAYPSAQNKFQPMGSSASQLRSS
```

Most Deleted  Most Deleted

---

**INFOLINK**

Sequence Polymorphism annotations are derived from the family tables derived from BLAST/FASTA output and are generated internally by GeneMine. Accordingly, linked information is not available with this annotation type when "About This Annotation" is clicked. Instead, a generalized help message will appear in the Information Window if the Infolink is invoked. To view the raw BLAST/FASTA information that GeneMine used to generate the information, return to
the initial results by using the left arrow to cycle back through cached hypernotes in the Information Window.

"About this Annotation" for known mutations brings up the protein file in the Information Window.

**STRUCTURAL HOMOLOGS**

Sequences found by sequence similarity search (see "Sequence Homologs"), by a BLAST search against the SCOP database. Sequence hits are clustered most typically by the SCOP superfamily classification system, so that sequences belonging to the same broad fold and functional class are clustered together. These hits may provide templates for modeling query sequences of unknown structure. SCOP is the "Structural Classification of Proteins" – a database and server in Cambridge, U.K. Click "About this annotation" to view statistical and other supporting data in the Information Window.

**DEFAULT DISPLAY STYLE**

Double line spanning region of homology, with text identifier of homolog: the structural superfamily name for the sequences in the cluster. Percent identity and number of proteins in the SCOP superfamily are indicated (e.g., 44% identity (11) coagulation factor IX).

Families comprised of a single sequence, or multiple sequences that have very close sequence identity (by default 95%) – SINGLE_GENE.IDENTITY_MIN=95% – are shown with double cyan lines. Families consisting of more diverse sets of sequences are shown with double white lines.
In the figure the top annotation is in cyan, indicating that it is a one-sequence cluster with 70% identity to the query sequence. (Cyan may also be used to indicate multi-sequence clusters when the sequences are very closely related.) The lower annotations show multi-sequence clusters with the number of sequences in each cluster, and the percent identity of the cluster leader to the query.

Where more than one sequence is present in a cluster, the percent identity is the average of all sequences in the cluster, not of the leader sequence.

**SCOP BACKGROUNDER**

The Structural Classification of Proteins (SCOP) database is maintained at the MRC Laboratory of Molecular Biology and Centre for Protein Engineering Cambridge, UK. The following summary is an edited excerpt from the SCOP web page.

The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known, including all entries in Brookhaven National Laboratory’s Protein Data Bank (PDB). It is available as a set of tightly linked hypertext documents which make the large database comprehensible and accessible. In addition, the hypertext pages offer a panoply of representations of proteins, including links to PDB entries, sequences, references, images and interactive display systems. World Wide Web URL http://scop.mrc-lmb.cam.ac.uk/scop/ is the entry point to the database.
Existing automatic sequence and structure comparison tools cannot identify all structural and evolutionary relationships between proteins. The scop classification of proteins has been constructed manually by visual inspection and comparison of structures, but with the assistance of tools to make the task manageable and help provide generality. The job is made more challenging--and theoretically daunting--by the fact that the entities being organized are not homogeneous: sometimes it makes more sense to organize by individual domains, and other times by whole multi-domain proteins.

**CLASSIFICATION**

Proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy, but the principal levels are family, superfamily and fold, described below. The exact position of boundaries between these levels are to some degree subjective. The SCOP evolutionary classification is generally conservative: where any doubt about relatedness exists, new divisions are added at the family and superfamily levels. Thus, some researchers may prefer to focus on the higher levels of the classification tree, where proteins with structural similarity are clustered.

The different major levels in the hierarchy are:

**Family**

Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. However, in some cases similar functions and structures provide definitive evidence of common descent in the absence of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%.

**Superfamily**

Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies. For example, actin, the ATPase domain of the heat shock protein, and hexokinase together form a superfamily.

**Fold**

Proteins are defined as having a common fold if they have same major secondary structures in same arrangement and with the same topological connections. Dif-
ferent proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin; the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

**INFOLINK**

Click "About this annotation" to view statistical and other supporting data in the Information Window. Structural Homology annotations are derived from the family tables derived from BLAST/FASTA output.

**CLOSELY RELATED HOMOLOG**

Generated by BLAST and FASTA searches against many databases, the same kind of cluster and family alignment — pairwise alignments and grouping into sub groups and families — is carried out with these annotation alignments as with sequence alignments. This annotation class represents sequences with very close identity to the query sequence. By default, the minimum percent identity required is 90%.

Within each family, GeneMine seeks to map a signature of internal conservation that can be a useful independent criterion for assessing whether the query sequence really matches that family. GeneMine attempts to answer the question: does the query sequence’s homology to the family mirror the family’s internal homology? GeneMine annotates the top 10% conserved residues within the family (purple + annotation). The maximum number of homology families allowed is 100, based on the observation that for some proteins there are many interesting families.

Regions of significant homology are annotated on the query sequence with labeled white lines, one per family. The family homologies are annotated slightly differently for results returned from SCOP BLAST than they are for Swiss FASTA. SCOP homologies are listed as "Structure (identity%): superfamily-name".

Superfamily-name is the name of the SCOP superfamily. (See the SCOP Cambridge, UK server). Identity is the overall average sequence identity of the family members vs. the query sequence.

Swiss FASTA homologies are listed as "identity% Identity: swissprot-name", where identity is the overall average sequence identity of the clustered homologs.
vs. the query sequence, and swissprot-name is the "description" name of the top-scoring Swiss Prot sequence in the cluster.

Click "About this annotation" to view a statistics table for the cluster in the Information window; follow a blue link to view the raw FASTA/BLAST results.

**DEFAULT DISPLAY STYLE**

Graphic: solid cyan bar, with text identifier of close homolog, including percent identity and number of proteins in closely homologous group (e.g., 71% identity (17): PROTEIN C PRECURSOR).

![Figure 76. Closely related homolog display style](image)

**EST EXPRESSION**

This annotation is calculated in a way similar to the Sequence Homology and Structural Homology annotations: a BLAST search is done against databases such as dbEST, and the families are clustered according to sequence similarity.

Protein sequences are searched against the dbEST database to find expression information. The search is run using BLASTX at NCBI, and clustered, processed as usual, with the usual virtual-document reports and links for loading or exploring the hits further.

Additionally, a dictionary of expression terms (e.g. "heart" "brain" "thymus" etc.) is searched against all the results, to elucidate patterns of tissue expression from the ESTs. The expression dictionary (expression.hot) is located in the Look v3/Discovery Engine install directory and is editable at the UNIX shell. The utility of this list is to extract and display a meaningful term, if present, from the EST file header information.
EST expression mapping on genomic DNA: using dbEST via BLASTN. Maps regions and frame of DNA that is expressed, or at least transcribed, with the same style of expression annotation applied to protein sequences, showing tissue localization, etc.

Discovery Engine has two clustering schemes for EST data:

- Clustering into gene families, with a 50% identity cut-off, the same default value as in other homology searches.
- Clustering by individual genes, with a 95% identity cut-off. This option can be made the default by editing the Worldbase configuration file, or can be applied directly by the user via the Sequence Queries dialog.

Access EST annotations from Sequence:Sequence Queries. Among the many available options in the pull-down are "dbEST Expression Annotations" for protein sequences, and "dbEST Transcription Annotations" for DNA sequences. The default cut-off for families is 95% identity.

**EST CLUSTERS**

GeneMine clusters ESTs into broad families to reduce redundancy and make functional patterns apparent (just as it does for protein homologies). However, users may want to see the ESTs clustered only at very high stringency (i.e., a cluster representing a single gene, not a family of related genes). These annotations are not turned on by default, but are accessed from the Sequence Queries dialog, "Expression, by gene" for protein sequences, "Transcription, by gene" for DNA sequences. The threshold cutoff for families is 95% identity, but can be adjusted within the Worldbase Perl scripts.
INFOLINK

Click "About this annotation" to view statistical and other supporting data in the Information Window. EST annotations are derived from the family tables derived from BLAST/Fasta output.

Default display style

Line spans region of homology to EST cluster. Multi-gene clusters are white; single member clusters are turquoise. Text identifier (% identity, number in family, tissue expression pattern, if present, otherwise dbEST name) accompanies line (e.g., 58% identity (2): yq58b08.s1 Homo sapiens cDNA clone 199959 3’ s).

Figure 78. EST expression homology display style

HOMOLOGY FINGERPRINT

Following BLAST and FASTA searches against many databases, patterns of sequence identities and deletions are tabulated across the total set of sequence homologs. These patterns can be displayed with the sequence homolog cluster "lines" ("Sequence Homologs") or by themselves to highlight key functional regions across sequence families.

This category of annotations can be split into Closely similar and Family Homology.

Closely Similar Homology

The Close Similarity Homology annotation marks the residues that are most conserved for all the homologs found by each type of search (e.g. SCOP, SwissProt,
A conservation score is calculated for each residue position of the query sequence. This score is the maximum percent identity of all homologs for which the amino acid at that position is not identical to the one in the query sequence. The top 10-25% of the residues with the highest conservation scores is marked as core homologies, the "fingerprint residues."

Being based on values calculated by Discovery Engine, annotation links through the "About This Annotation..." command are not available for core homologies.

**Default display style, Most Conserved**

Cyan character for each conserved residue, in linear arrays. Each line of characters represents a homology cluster.

**Family Homology**

The Family Homology annotation is similar to the Most Conserved annotation, except that instead of being calculated for all homologs, it is calculated for each family. Because there are fewer sequences in the calculation, the statistics are coarser, and the top 10-50% of the residues with the highest conservation scores is marked.

This annotation is only calculated if the sequence identity between the query sequence and the family is less than a default threshold of 70%. Default display style, family homology

Purple character for each conserved residue, in linear arrays. Each line represents a homology cluster.

**CELLULAR TUMOR ANTIGEN P53 (PHOSPHOPROTEIN P53).**

![Figure 79.homology fingerprint display style](image)
VALIDATION OF METHOD

A study conducted at MAG to confirm the accuracy of calculated annotations was directed at the genomes of M. jannaschii and H. influenzae. An assortment of ORFs in the genomes of these organisms were identified by TIGR (Gilbert et al. Science 276:1724-25, Jun 13; 1997). These ORFs were retrieved from the WWW and read into GeneMine/Discovery Engine. Discovery Engine was then asked to perform regular workup level annotations for the ORFs called M2, M5, and M9 in the Methanococcus jannaschii genome. In each of these cases, it confirmed the findings of the article and provided even more information. Snapshots of the annotations are shown below the table.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Protein</th>
<th>Result</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>M01</td>
<td>30S Ribosomal protein S14</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M02</td>
<td>Yqgp protein</td>
<td>CONFIRMED</td>
<td>showed homologies to membrane proteins</td>
</tr>
<tr>
<td>M03</td>
<td>Amidophosphoribosyltransferase</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M04</td>
<td>Unknown</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>M05</td>
<td>Asparaginesynthetase</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M06</td>
<td>Modification methylase</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M07</td>
<td>Modification methylase HINCII</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M08</td>
<td>Helicase</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M09</td>
<td>Unknown</td>
<td>N/A</td>
<td>found possible homologies to proteins: interferon-gamma binding protein xeroderma pigmentosum group C complementing factor</td>
</tr>
</tbody>
</table>

Table 4. M jannaschii ORFs, by TIGR

METHANOCOCCUS JANNASCHII GENOME

Shown below are snapshots of the Look v3 annotations on the ORFs in question from the Science article "Dealing with Database Explosion: A Cautionary Note." Sequences were obtained at the URL referenced in the article:

The ORFs called M2, M5, and M9 in the article were retrieved from the web and inserted into Look v3. The program was then directed to perform regular annotations for each of the sequences to see what it would discover. In each case it confirmed the findings of the article and provided even more information.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Protein</th>
<th>Result</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>M01</td>
<td>30S Ribosomal protein S14</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M02</td>
<td>Vopp protein</td>
<td>CONFIRMED</td>
<td>showed homologies to membrane proteins</td>
</tr>
<tr>
<td>M03</td>
<td>Amido phosphoribosyltransferase</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M04</td>
<td>Unknown</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>M05</td>
<td>Asparaginesynthetase</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M06</td>
<td>Modification methylase</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M07</td>
<td>Modification methylase HINCII</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M08</td>
<td>Helicase</td>
<td>CONFIRMED</td>
<td></td>
</tr>
<tr>
<td>M09</td>
<td>Unknown</td>
<td>N/A</td>
<td>found possible homologies to proteins: interferon-gamma binding protein xeroderma pigmentosum group C complementing factor</td>
</tr>
</tbody>
</table>

Table 5. M. jannaschii ORFs annotated by Discovery Engine
Figure 80. Membrane protein similarity annotations
Figure 81. Asparagine synthetase-like region annotation

Figure 82. Possible homology to X. pigmentosa and interferon
**HAEMOPHILUS INFLUENZAE GENOME**

For a second validation exercise, the entire Haemophilus influenzae genome was downloaded from the TIGR database at ftp.tigr.org. The protein translations of several predicted genes provided by the database were inserted into GeneMine and a regular workup was launched. Genes that encoded hypothetical and characterized proteins or had putative function were used.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Interesting Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>HI004</td>
<td>hypothetical protein</td>
<td>found the DNA sequence (100% identity)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>core homology and family homology annotations are displayed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>found homology (33% identical) to a diacylglycerol kinase, possible function?</td>
</tr>
<tr>
<td>HI0067</td>
<td>DNA mismatch repair protein</td>
<td>found many mismatch repair proteins and their homologs</td>
</tr>
<tr>
<td>HI0092</td>
<td>hypothetical protein</td>
<td>found DNA sequence (100% identity)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>found weak homology (25% identity) with a gluconate permease</td>
</tr>
<tr>
<td></td>
<td></td>
<td>found homology with other hypothetical proteins</td>
</tr>
<tr>
<td>HI0163</td>
<td>putative murein gene regulator</td>
<td>found DNA sequence (100%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>found some other hypothetical proteins</td>
</tr>
<tr>
<td></td>
<td></td>
<td>found homology to a cellulose-binding protein</td>
</tr>
<tr>
<td>HI0556</td>
<td>putative glucose-6-phosphate dehydrogenase isozyme</td>
<td>found DNA sequence (100% identity)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>found homology to other glucose-6-phosphate 1-dehydrogenase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>found homology to a putative glucose-6-phosphate</td>
</tr>
<tr>
<td>HI1277</td>
<td>putative ATPase</td>
<td>recognized an ATP-binding site motif</td>
</tr>
<tr>
<td></td>
<td></td>
<td>found structural homology to CoA-dependent acetyltransferases</td>
</tr>
</tbody>
</table>

Table 6. Discovery Engine annotations of H. influenzae
Figure 83. Hypothetical protein

Figure 84. DNA mismatch repair protein
Figure 85. Putative ATP-ase annotation
6. STANDARD ANNOTATIONS

CHAPTER CONTENTS

- Sequence databases annotations
- Motif Annotations
- SSP annotations
- Domain Annotations
- Active site
- Specificity Pocket
- Epitope
- Post-translational modifications
- Metabolic pathway/enzyme activity
- Genetic map/linkage
- Disease Association
- Restriction Sites
- Open reading frame
- PCR primer prediction
- Disulfide bridges
- Literature
SeqDBAnnotation

Standard annotations differ fundamentally from the calculated annotations described in the previous chapter in that no client-side processing is done on them. Standard annotations are returned as a SeqDBAnnotation type.

**DOCUMENT-TYPE: SeqDBAnnotation**

This annotation reader uses extremely simple parsing. The text should begin with a line:

```
ANNOTATION-FORMAT
```

A series of lines follows giving a single statement per line:

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEATURE=feature-type</td>
<td>This field specifies the type of feature observed: Disulfide, Active, Product, Domain, strand, helix, turn, variant, conflict, Modified, Motif, Binding, MAP-LOCATION, LITERATURE, ORF+, ORF-, PRIMER+, PRIMER-, RESTRICTION, Peptide, Signal, Propep, Metal, Varsplic, mutagen, Lipid, Thiolest, Thioleth, Misc_site, Transit, Ca_bind, Dna_bind, Np_bind, Transmem, Zn_fing, Similar, Repeat, EC</td>
</tr>
<tr>
<td>LABEL=label-text</td>
<td>This is the label shown on the annotation. Any desired label-text, including whitespace, may be provided, up to the end of the text line.</td>
</tr>
</tbody>
</table>
Table 7. SeqDBAnnotation formats

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
</table>
| INFOLINK     | This is an optional field that can specify a worldbase path to link with this annotation. This variable specifies annotation document hyperlink paths to link to this annotation as a standard Worldbase path. Can include whitespace and title information (#t=title info).
| E.g., INFOLINK=||prosite-id:PS500011#t=Motif:GLU_CARBOXYLATION is a Worldbase path to the prosite-id service with title=Motif:GLU_CARBOXYLATION |
| RANGE        | Annotation Range specifies the numerical range of sequence positions to link to this annotation. |
| SCORE        | This specifies the log-odds score for these annotations. |
| END-ANNOTATION | Each annotation must end with this line |

**MOTIF ANNOTATIONS**

**FEATURE=Motif**

Sequence patterns that match characteristic protein features are often not detected by sequence similarity searches. Functional motifs are predictive, being based on matches between the query sequence and the PROSITE database, or other pattern search methods such as PFAM, to identify potential functional features. Common matches that are likely to be false positives, such as N-glycosylation, found in most known protein sequences, are omitted.

Click "About this annotation" to view statistical and other supporting data in the Information window.
Discovery Engine Reference Manual

**DEFAULT DISPLAY STYLE**
Green line spanning motif region, with text identifier (e.g., motif Scorpion short toxins signature).

**SECONDARY STRUCTURE PREDICTION (SSP) ANNOTATIONS**

FEATURE=strand
FEATURE=helix
FEATURE=turn

**EXAMPLE**
Helix, turn, sheet, or random coil (i.e., no secondary structure).

**DEFAULT DISPLAY STYLE**

<table>
<thead>
<tr>
<th>SwissProt, etc.: Predicted</th>
<th>Helix</th>
<th>Strand (Sheet)</th>
<th>Turn</th>
<th>Random Coil</th>
<th>Type style</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Open box</td>
<td>Open arrow indicating N-to-C directionality</td>
<td><em>=</em> symbol (SwissProt only) above each residue</td>
<td>—</td>
<td>italics, to indicate uncertainty</td>
</tr>
</tbody>
</table>
| PDB: Observed, recorded by author | Solid cylinder with black dot at N-terminal end | Solid arrow indicating N-to-C directionality | *
|                            | solid line | plain type |

Table 8. SSP display modes

Each sequence in an alignment has its own color for secondary structure symbols. Colors are allocated from white to red, in the order in which they appear in the color palette.

For PDB records which do not contain secondary structure information, a solid white line running the length of the sequence is displayed.
The Secondary Structure Prediction Program (SSP) predicts a-helix and b-strand segments of globular proteins. The server is at the Department of Cell Biology, Baylor College of Medicine (BCM).

The segment-oriented method is designed to locate secondary structure elements and uses linear discriminant analysis to assign segments of a given amino acid sequence to a particular type of secondary structure, by taking into account the amino acid composition of internal parts of segments as well as their terminal and adjacent regions. Four linear discriminant functions were constructed for recognition of short and long a-helix and b-strand segments, respectively. These functions combine 3 characteristics: hydrophobic moment, segment singlet and pair preferences to an a-helix or b-strand. To improve the prediction accuracy of the method, a simple version which treats multiple sequence alignments that are used as input in place of single sequences has been developed.

**Accuracy of SSP**

Adapted from the PSSP web page at Baylor.

Overall 3-states (a, b, c) prediction gives ~65.1% correctly predicted residues on 126 non-homologous proteins using the jack-knife test procedure (The accuracy is good if you have no homologous sequences to apply Sander et al. method (Rost,Sander, Mol.Biol,1993,232,584-599) that has about 71% accuracy with using these sequences and about 61% without them). Analysis of the prediction results shows a high prediction accuracy of long secondary structure segments (~89% of a-helices of length greater than 8 and ~71% of b-strands of length greater than 6 are correctly located with probability of correct prediction 0.82 and 0.78 respectively). Using the mean values of discriminant functions over the aligned sequences of homologous proteins, we achieved a prediction accuracy of 68.2%. It must be mentioned that our variant of nearest-neighbor algorithm with using multiply sequence alignments of homologous proteins has 72% accuracy and 67.6% accuracy without homologous proteins. (see "nnssp" program of this server). Reference


Solovyev V.V., Salamov A.A. Predicting a-helix and b-strand segments of globular proteins. CABIOS (1994), V.10,6,661-669
**Discovery Engine Reference Manual**

**DOMAIN ANNOTATIONS**

FEATURE=Domain

Region (range of sequence positions) known or predicted to represent a functional domain or region. Domain annotations may be taken from feature table databases (e.g., SwissProt, GenBank, PIR), domain predictions (e.g., PRODOM), or other predictions such as transmembrane segment.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

**Examples**

Signal peptide, pro-peptide, ligand-binding domain, trypsin-sensitive domain, transmembrane region.

**Default display style**

Solid white line spans domain region; text identifier (e.g., PROTEIN C HEAVY CHAIN domain).

**PRODOM SERVER**

The ProDom database is a comprehensive collection of protein families. It was constructed by clustering all complete protein sequences in SwissProt by the DOMAINER algorithm. The novelty of ProDom is that the modular arrangement of proteins has been taken into account and whenever domain boundaries were detected the sequences were cut to produce consistent families of domains.

**REFERENCE**


**ACTIVE SITE**

FEATURE=Active
Sequence (residue) positions in enzymes which are known or predicted to be involved in catalytic activity of a protein. Taken from feature table databases (SwissProt).

Click "About this annotation" to view statistical and other supporting data in the Information Window.

Example

Serine protease catalytic triad (His/Asp/Ser).

Default display style

Red * at active site position(s); text identifier (e.g., ACT_SITE CHARGE RELAY SYSTEM).

SPECIFICITY POCKET

FEATURE=Binding

Include binding sites and prosthetic groups, from feature table database information in SwissProt.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

Examples

Heme, flavin, etc.

Default display style

Orange character at site, with text identifier (e.g., BINDING HEME (COVALENT)).

EPITOPE

(Place-holder in this version of Discovery Engine).
POST-TRANSLATIONAL MODIFICATIONS

FEATURE=Modified

Many proteins have amino acids which are modified chemically after the protein is synthesized. This information, when known, is sometimes available in the feature table database. Discovery Engine extracts this information and shows it graphically against the sequence as for other annotations. Click "About this annotation" to view statistical and other supporting data in the Information Window.

Examples

Glycosylation (N- or O-), amidation, hydroxyproline.

Default display style

Green ▲ character at modification site, with text identifier (e.g., MOD_RES HYDROXYLATION).

METABOLIC PATHWAY/ENZYME ACTIVITY

FEATURE=EC

When an EC (Enzyme Commission) number is present in the record header, metabolic pathway information is retrieved from the ExPASy ENZYME Database.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

Example

Lysyl endopeptidase, from Achromobacter lyticus.

Default display style

White line with text identifier (e.g., EC Classification 3.4.21.50) Metabolic pathway/enzyme activity.
**GENETIC MAP/LINKAGE**

**FEATURE=MAP-LOCATION**

A range of sequence positions matched against an STS. Connects sequence to chromosomal maps – done by sequence similarity searches against the STS database. Similar to ESTs, STSs are "Sequence-Tagged Sites" – small regions of genomic DNA that have been cloned and sequenced from a known chromosomal location.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

**Default display style**

Blue line spans match region, accompanied by text identifier (e.g., STS 21899 on Homo sapiens Chromosome 2).

**DISEASE ASSOCIATION**

Disease associations are defined in the diseases.hot dictionary.

Disease association terms (e.g. carcinoma, diabet, etc.) are searched against all the results to elucidate patterns of expression from the ESTs. The disease association dictionary (diseases.hot) is located in the Look/Discovery Engine install directory and is editable at the UNIX shell.

The disease association dictionary is cross-category and could include hits categorized as Domain, Homology and Motif annotations. Disease associations will be displayed when any of the relevant annotation categories are toggled on. Conversely, when Disease annotation is turned on, all other annotation categories (and their icons) in which the disease term is found will also be turned on.

Click "About this annotation" to view STS record showing experimental source, procedure, detailed mapping information, local homologies, and other supporting data in the Information Window.

**Default display style**

As for the annotation type whence the disease association originated (Domain, Homology, Motif, etc.).
**RESTRICTION SITES**

**FEATURE=restriction**

(DNA annotation only) Shows unique restriction sites, indicating enzyme recognition sites that occur only once in the DNA segment. Performed using the "WebCutter" program through the Baylor College of Medicine site. The range of nucleotides constituting the recognition site is shown.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

**Default display style**

Bracketed white line at cut site, with name of enzyme (e.g., HindIII).

![Figure 86. Display style for Restriction Sites](image)

**OPEN READING FRAME**

**FEATURE=ORF+**

**FEATURE=ORF-**

(DNA annotation only). ORF prediction is done by NCBI’s Gene Finder server. ORFs are detected by TBLASTN homology search against the NCBI nr protein dataset. Hits are clustered as for protein, linked to
virtual-documents reporting the search results in tabular form for each family, with links for loading any of the sequences. Annotations show the exact placement, frame, and direction of the ORFs.

The translated amino acid sequence may be loaded using the annotation’s infolink. Note that amino acid sequences coded by the complementary (antisense) strand are placed at the 3’ end of the DNA sequence, not adjacent to the coding region. This is to avoid displaying amino acid sequence in the C → N direction, which would be necessary, because Discovery Engine displays DNA sequence in the single-stranded sense orientation only.

Prediction of regions in DNA which might be translated into protein sequences. For cDNA sequences, these should be contiguous. For genomic DNA, introns of varying size (length) and number may be present between the exons, complicating the analysis of potential open reading frames. Discovery Engine by default uses the NCBI “GORF” server for this category.

Discovery Engine also performs a six-frame translation search against protein sequence databases to discover potential open reading frames by homology. Results of this search, including links to load the inferred amino acid translation, are shown as homology annotations.

Click "About this annotation" to view in the Information Window the translation, and the link to load it for workup, alignment and analysis.

Note on display of proteins translated from complementary strand in opposite direction.

**Default display style**

White line with arrowhead to show directionality of ORF. Text label indicates frame number.

**PCR PRIMER PREDICTION**

FEATURE=PRIMER+

FEATURE=PRIMER-

(DNA annotation only). Displays ranges of sequences which might be useful for creating synthetic DNA primers for PCR to clone out DNA fragments from larger stretches of DNA. Displays location and orientation (arrow) of the predicted possible primer sequences. Obtained by
sending sequence to Baylor College of Medicine site which uses the xprimer program.

Click "About this annotation" to view statistical and other supporting data in the Information Window.

PCR primer prediction is returned by the primertx program at Baylor College of Medicine (BCM) at:

http://dot.imgen.bcm.tmc.edu

The BCM Search Launcher references the Alces site in Minnesota at:

http://alces.med.umn.edu/xprimerinfo.html

The xprimer program outputs in both text and graphics versions. primertx modifies the output by combining the results in a text/tabular display.

**Default display style**

Inward-pointing arrows at forward and backward primer sites with set number and score.

**xprimer background**

xprimer is designed to select of sets of primers along very large queries which may be permeated with repeat regions. xprimer attempts to filter out the repeat regions by comparing target regions against a database of known octanucleotide repeats. Primers are all calculated in a relatively narrow Tm range. xprimer is also useful in more traditional PCR applications.
Discovery Engine Reference Manual

Scoring

The score reported in GeneMine is the frequency of the primer's terminal octanucleotide in the model database. Higher model scores mean that a primer is located in a repeat-containing region. The primers are sorted with lowest model scores being at top.

xprimer also calculates another score, the local score, which is the number of times the primer occurs in the query. This score is not reported by GeneMine.

The repeat databases HR and BR3X contain human repeated sequences. Human, yeast, and nematode genome models are available.

Parameter adjustment

There are five adjustable parameters. GeneMine reports the primers as calculated by the server defaults:

- Repeat database: BR3X
- Genome Model: HUMAN.pro
- Number to show (each direction): 12
- Tm range: 63.0-64.0
- Size range: 21-27

To calculate primers with non-default values, open Netscape and go to the server and adjust the parameters as desired. Then return to GeneMine and use the Open from Netscape command. The results will appear in the Information window. Note that only primers calculated using default values will appear as annotations attached to the sequence in the Sequence window.

xprimer Algorithm

The frequency occurrences of K-tuple (overlapping sequences of defined length, K) were computed from known human genome sequences. The significance of these frequencies for the whole human genome was tested by polymerase chain reaction (PCR). A computer program based on these results was written to choose primers to amplify DNA target sequences, either of human genes or of human infectious agents. The software also gave nested primer sequences which were used to synthesize non radioactive probes by PCR.
Discovery Engine Reference Manual

Reference


DISULFIDE BRIDGES

FEATURE=Disulfide

Often a key structural element from feature table databases.

Click "About this annotation" to view information extracted from the feature table of the sequence record in the Information Window.

Default display style

Yellow dot below cysteine residue, with text label "S-S bridge," numbered in the order of occurrence in the source record’s Feature Table.

LITERATURE

FEATURE=LITERATURE

Comprehensively lists all annotations in the current session, including literature links as well as links pointed to by feature table expressions.

Feature table databases frequently reference literature links such as Medline and OMIM. Click "About this Annotation” to view abstracts and further hyperlinks to documents and sequences.

A second literature link is limited to bibliographic citations. When these are present in a search result, the icon will appear among the sequence icon cluster in the Sequence Window. Typically these links lead to Entrez Pubmed, but may also lead to OMIM and other databases.

Note that the Literature icon does not appear on the Annotation Toolbar.
Default display style

White line (normally spanning the entire sequence) with bibliographic citation.
7. PRIMARY SOURCES

INTRODUCTION

IMPORTANCE OF Worldbase

This chapter describes at a high level the information content of the four primary databases: Brookhaven National Lab’s PDB protein structure database; the PIR 2-D protein database; GenBank; and the SwissProt protein database.

Chapter sections include Worldbase records (excerpts from the worldbase.src file) to illustrate how Discovery Engine accomplishes particular tasks of information retrieval from the primary information sources. When reviewing the records presented here, refer back to Chapter 4, "Worldbase," as a primary information source on Worldbase organization and definitions.

The Worldbase records will give the Discovery Engine user a perspective on the possibilities for customized information retrieval. Although many of the fields present in the source database records are not called by the default Worldbase file, these fields could be incorporated into custom Worldbase records so that searches would follow particular routes and amplify the results of a given line of inquiry. For example, a researcher might want to search by ATCC record number, or by author name, or by some other identifier not named in the Worldbase defaults. To do so s/he would first identify the data type and its format requirements (e.g., column numbers) and then write a custom script in Worldbase, following the guidelines set forth in Chapter 3. After testing to guarantee that the script worked properly, it would be incorporated into the user’s personal Worldbase file to select the information it identified.

Note that this chapter does not include the detailed and exhaustive information necessary, such as column numbers and special characters. To obtain current information about the format characters in a given database, refer to the current documentation at the provider’s web site.
CAVEAT ON PATH= VARIABLE

Many of this chapter’s descriptive sections about Discovery Engine’s information sources are followed by sample scripts extracted from the worldbase.src file. These scripts are included to illustrate how Worldbase can be configured to access these information sources. But users should be aware that by their nature network sources are dynamic and changeable, and that the syntax indicated here is liable to produce errors or unanticipated results if changes in the source’s server address and syntax are not incorporated into Worldbase revisions.

PRIMARY DATABASES

Most of the information that GeneMine accesses ultimately originates in the four sequence databases: PDB, PIR, SwissProt, GenBank.

PDB

The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules. It contains atomic coordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic structure factors and NMR experimental data.


HOW Worldbase ACCESSES PDB

Worldbase RECORD FOR PDB SEARCH AT MAG

# MAG PDB Server
SOURCE=mag-pdb
HISTORY=pdb-keyword:%key%
NAME=PDB Search at MAG
Discovery Engine Reference Manual

PATH=http://pdb.mag.com/cgi-bin/pdb_search?
SYNOPSIS=Search for PDBs by keyword (i.e. "hiv* and protease") or by ID (i.e. "1enh"). Keywords must be three characters or longer.
TITLE=MAG PDB Search on %key%
WHITESPACE=+
WILDCARD=* 

Worldbase RECORD FOR PDB AT BROOKHAVEN 
SOURCE=brookhaven-pdb
HISTORY=pdb:%key%
LINEWIDTH=120
NAME=Brookhaven PDB SERVER
OUTPUT=Mol
PATH=@@@http://www.pdb.bnl.gov/cgi-bin/send-pdb?filename=%key%
SERVICE=pdb&
SYNOPSIS=Search for PDB structure files at Brookhaven (main PDB) by keyword (i.e. "hiv* and protease") or by ID (i.e. "1enh").
TITLE=PDB Entry %key%
UI=hide

Worldbase RECORD FOR PDB AT NIH 
# NIH 
SOURCE=nih-pdb-key
HISTORY=pdb-keyword:%key%
NAME=PDB Search at Brookhaven
OR=+
PATH=http://molbio.info.nih.gov/cgi-bin/pdb?%key%
SERVICE=pdb-keyword&
SYNOPSIS=Search for PDBs by keyword (i.e. "hiv and protease") or by ID (i.e. "1enh").Supports AND and OR, but no wildcards (*).
TITLE=PDB Search on %key%
WHITESPACE=+

**PRERELEASE PDB**

This server allows you to search the contents of the Brookhaven PDB sites, which include unreleased entries. This server can be searched as described above in the Protein Data Bank. In order to open a PDB file retrieved from the prerelease PDB server, you will have to check the List All Files radio button in the File-->Open dialog.

**Worldbase RECORD FOR PRERELEASE PDB**

SOURCE=brookhaven-pdb-info
HISTORY=pdb-info:%key%
NAME=Prerelease PDB SERVER
PATH=http://www.pdb.bnl.gov/cgi-bin/send-pdb?filename=%key%&short=1
SERVICE=pdb-info
STRIP=AZ
TITLE=Info on PDB Entry %key%
UI=hide

**SCOP, A DERIVATIVE OF PDB**

The Structural Classification of Protein database, maintained at the Laboratory of Molecular Biology at Cambridge, UK, is derived from the Protein Database. As of March, 1997 it contained 5493 PDB entries comprising 10781 Domains and 138 Literature References.

The SCOP database, created by manual inspection and a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. It provides a broad survey of all known protein folds, detailed informa-
tion about the close relatives of any particular protein, and a framework for future research.


Worldbase RECORD FOR SCOP

# STRUCTURAL CLASSIFICATION OF PROTEINS
SOURCE=scop-key-ui
HISTORY=scop-key:%key%
NAME=Structural Classification of Proteins
PATH=||scop-key:%key%
SERVICE=pdb-keyword&
SYNOPSIS=Search for PDBs by keyword. Use ONLY ONE keyword (i.e. homeo*).
TITLE=SCOP Search on %key%

PIR

PIR, the Protein Information Resource, is maintained at Brookhaven National Lab. It is a primary source both for alignments and sequence annotations.

Worldbase RECORD FOR PIR HOMOLOGY ALIGNMENTS

# PIR SEQUENCE SET
SOURCE=eerie-pir-fasta
FILTER=a"$LOOK_DIR/exec/eerie-pir.awk"
HISTORY=pir-fasta
INPUT=Seq
NAME=FASTA PIR Homology Annotations EERIE
OUTPUT=BLASTHomology
PATH=@http://vega.crbl.cnrs-mop.fr/bin/nph-fasta_query.pl?br=no_record&ou=www&ad=&db=pir+%pr=fasta&ml=&kt=1&ma=BLOSUM62&sc=200&al=200&ot=on&co=seq,&qu=<<<%seq%>>>&ty=Protein
STRIP=?atgcu
**Worldbase RECORD FOR PIR ANNOTATIONS**

# PIR ANNOTATION SOURCE
SOURCE=pir-annot-auto
INPUT=Seq
LINEWIDTH=160
NAME=PIR Annotations
OUTPUT=SeqDBAnnotation
PATH=@@http://www.embl-heidelberg.de/srs/srsc?[PIR-id:<<<%s>>>
SERVICE=pir-annot-auto
SYNOPSIS=Find known sequence features for a PIR entry (via SRS at EMBL).
TITLE=Feature Search on <<<%s>>>

**SWISSPROT**

SWISS-PROT contains sequences translated from the EMBL Nucleotide Sequence Database, prepared by the European Bioinformatics Institute. A small part of the information in SWISS-PROT was originally adapted from the Protein Sequence Database of the Protein Information Resource (PIR) supported by the Division of Research Resources of the NIH, National Biomedical Research Foundation, Georgetown University Medical Center.

**How Worldbase ACCESSES SWISS-PROT ANNOTATIONS**

**Worldbase RECORD FOR SWISS-PROT ANNOTATIONS (MAIN)**

Find known sequence features for a SwissProt entry (via SRS at IUBIO).

# SWISSPROT ANNOTATION SOURCE
SOURCE=iubio-swiss-annot
HISTORY=swiss-annot
INPUT=Seq
Worldbase record for Swiss-Prot annotations (Backup 1)
Find known sequence features for a SwissProt entry (via SRS at NYU).

Worldbase record for Swiss-Prot annotations (Backup 2)
Find known sequence features for a SwissProt entry (via SRS at Sanger Centre, UK).
**Discovery Engine Reference Manual**

NAME=Swiss-Prot Annotations (Backup 2)
OUTPUT=SeqDBAnnotation
PATH=@@http://www.sanger.ac.uk:80/srs/srsc?[SWISSPROT-id:<<<%%s>>>]+-sf+EMBL
SERVICE=swiss-annot&
SYNOPSIS=Find known sequence features for a SwissProt entry (via SRS at Sanger Centre, UK).
TITLE=Feature Search on <<<%%s>>> 

**Worldbase record for Swiss-Prot Annotations (Backup 3)**
Find known sequence features for a SwissProt entry (via ExPASy server).
SOURCE=expasy-swiss-annot
HISTORY=swiss-annot
INPUT=Seq
LINEWIDTH=120
NAME=Swiss-Prot Annotations (Backup 3)
OUTPUT=SeqDBAnnotation
PATH=@@http://expasy.hcuge.ch/cgi-bin/get-sprot-entry?<<<%%s>>> 
SERVICE=swiss-annot&
SYNOPSIS=Find known sequence features for a SwissProt entry (via ExPASy server).
TITLE=Feature Search on <<<%%s>>> 

**GENBANK**

Annotation information is derived from the GenBank Feature Table and Key-Word lines.

**How Worldbase accesses GENBANK**

**Worldbase record for ENTREZ GENBANK**

# entrez nucleotide sequence server
SOURCE=entrez-genbank-dna
HISTORY=genbank-dna:%key%
NAME=Genbank DNA from Entrez
OUTPUT=Seq
SERVICE=genbank-dna,entrez-dna
TITLE=Genbank DNA %key% from Entrez
UI=hide

**WORLDBASE RECORD FOR GENBANK (ALTERNATE SERVER)**

# GenBank Alternate (Japan)
SOURCE=DDBJ-gmp
NAME=GenBank (Alternate Server)
PATH=gopher://ftp2.ddbj.nig.ac.jp:70/7waissrc%3A/search/gb?%key%
SYNOPSIS=Alternative source for keyword searches of GenBank (DDBJ, Japan). Supports Booleans and wildcards (i.e. "homeo* and pou").
WILDCARD=*  

**WORLDBASE RECORD FOR GENBANK (IUBIO)**

# GDB GENBANK SERVICE
SOURCE=GenBank
NAME=GenBank (IUBIO)
PATH=gopher://ftp.bio.indiana.edu/77/.bin/genbankq?%key%
SYNOPSIS=Find DNA sequences from GenBank using the IUBIO (Indiana) server (i.e. "homeo* and xenopus").
TITLE=GenBank Search on %key%
WHITESPACE=+
WILDCARD=*
8. COLLATERAL SOURCES

INTRODUCTION

Typically, an annotation that appears in the Discovery Engine Sequence Window is derived from information retrieved by collateral searching across multiple databases. In this chapter we look at the information that Discovery Engine retrieves and processes from a categorical perspective, rather than from a server-based perspective, as in the previous chapter.

In this chapter the information sources that Discovery Engine draws on are first grouped in three superset categories:

- Homology information (e.g., BLAST on PIR)
- Feature information (e.g., Prosite motifs)
- Literature information (e.g., Medline references)

Within each superset category, annotations are listed alphabetically as they appear in the File:New:Keyword Search... and Sequence:Sequence Queries... drop-down lists.

CAVEAT ON PATH= VARIABLE

Many of this chapter’s descriptive sections about Discovery Engine’s information sources are followed by sample scripts extracted from the worldbase.src file. These scripts are included to illustrate how Worldbase can be configured to access these information sources. But users should be aware that by their nature network sources are dynamic and changeable, and that the syntax indicated here is liable to produce errors or unanticipated results if the source’s server address and syntax change.
**COMBINED INFORMATION SOURCES**

**ENTREZ SOURCES**

Entrez is a comprehensive set of links maintained by NCBI. Results available from Entrez include:

- GenPept Reports
- Sequence Reports
- FASTA Reports
- ASN.1 Reports
- MEDLINE Links
- Protein neighbors
- Nucleotide links

Netscape search

**GENQUEST SEARCH (Netscape)**

Sometimes you want to search using non-default parameter settings. To change a server’s search or stringency matching parameters, go to the server’s site. Copy/paste the query sequence into the Netscape form, change the settings as desired on the server’s page, and launch the search. Then return to the Discovery Engine UI and use the File:Open from Netscape command to import the result.

You can start Netscape automatically with the GenQuest search (Netscape) in the Sequence Queries drop-down list.

**OWL**

OWL is at:

http://www.bis.med.jhmi.edu/Dan/proteins/owl.html

The OWL database is a non-redundant protein sequence database produced from the following source databases:
SWISSPROT (Bairoch, A. & Boeckman, B. (1991))

- PIR(1-3) (Sidman, K., George, D., Barker, W. * Hunt, L. (1988))
- GenBank translations (Benson, D, Lipman, D., & Ostell, J. (1993))

No two sequences in this database are exactly the same and no two sequences show only "trivial" differences. The Web version derived from OWL has hot links to the following Databases, many of which have links among themselves and to other databases:

- PIR(1-3) - The Protein Identification Resource
- GenBank - DNA Sequence Database
- EMBL - The European Molecular Biology DNA Sequence Database
- DDBJ - The DNA Data Bank of Japan.
- PDB - The Protein Databank (3D structures)
- EC-Enzyme - The EC Enzyme Classification Database
- PROSITE - A Dictionary of Protein Sites and Patterns
- OMIM - Online Mendelain Inheritance in Man
- SWISS-2DPAGE - Two-dimensional Polyacrylamide Gel Electrophoresis Database
- REBASE - The Restriction Enzyme Database
- Refbase - A Protein Sequence Citation Database

**SEQUENCE DATABASES**

This source accesses several databases which are searched at the same time: PDB, PIR, SWISS PROT, and Genbank. Note that sequences and structures found in the PDB database must be downloaded as a separate operation from the Protein Data Bank server. These entries represent structures (molecules).
Worldbase RECORDS

Worldbase RECORD FOR GENQUEST SEARCH (NETSCAPE)

# PERFORM A SEARCH WITH NETSCAPE FORM
# ! in VIEWER prevents caching, forces relaunch on each Open
SOURCE=Genquest-NS
INPUT=Seq
NAME=GenQuest Search (Netscape)
PATH=|genquest-netscape.sh '<<<%seq%s>>>'
STRIP=[?]
SYNOPSIS=Starts Netscape with GenQuest HTML form, with your selected protein sequence loaded. Adjust the form parameters, and submit. When the results are returned, import them by using "File-> Open from Netscape."
TITLE=Netscape-Genquest search on <<<%%%s>>>
VIEWER=document!

Worldbase RECORDS FOR OWL

SOURCE=owl-jhu-get
LINEWIDTH=120
NAME=OWL Sequence Server
OUTPUT=Seq
PATH=@http://www.gdb.org/bin/bio/wais_q-bio?object_class_key=35&jhu_id=%key%
SERVICE=owl-id
TITLE=OWL Sequence %key%
UI=hide

# Filter trick takes the 1st owlref URL and forces it to be read as @http://...
SOURCE=owl-jh-annot
FILTER=P"%@9$s"="owlref"
INPUT=Seq

160
NAME=OWL Annotation Server
OUTPUT=SeqDBAnnotation
PATH=http://www.gdb.org/bio/search/FILT/owlsref.html?<<<%%s>>>
SERVICE=owl-annot-auto
TITLE=OWL Annotations for <<<%%s>>>
UI=hide

Worldbase RECORD FOR SEQUENCE DATABASES
# NIH Sequence database search, non-redundant PIR/Swiss/GenBank entries
SOURCE=nih-seq
NAME=Sequence Databases
PATH=gopher://gopher.nih.gov/7mindex:/molbio/all.mindex?%key%
SYNOPSIS=Keyword search of NIH gopher, non-redundant sequence databases
(GenBank, PIR, SWISSPROT). Supports Booleans and wildcards (i.e.
"insulin and pancrea*")
TITLE=NIH Sequence Search on %key%
WHITESPACE=+
WILDCARD=* 

HOMOLOGY INFORMATION

BLAST SEARCHES

The Basic Local Alignment Search Tool finds similar proteins in the database
and returns a list from which you can choose which sequence(s) to download. 
Discovery Engine sends your DNA or protein sequence to the BLAST programs
to search for homologous nucleotide or amino acid sequences. You will get a list
of related sequences and links to sequence and structure databases (pir, BDB, En-
trez, and GenPept.) You may use the links to retrieve sequences.

Note that PDB files must be retrieved from the Protein Data Bank. PDB files do
not contain sequences in a standard format. The results are returned in order of
p-value, with the default cutoff value as set by the BLAST program.
**BLAST2**

GeneMine supports BLAST2 database searching. Gapped alignments (i.e., alignments containing deletions and insertions) are produced, with potentially multiple regions of similarity being found between each pair of sequences. With a gapped alignment tool, homologous domains do not have to be broken into several segments. Also, the scoring of gapped results tends to be more biologically meaningful than ungapped results. The gapped alignment routines are integral to the database search itself, not a post-processing step grafted onto BLAST1 and thus yield better sensitivity. BLAST2 executes about 10% slower than its version 1.4 counterpart, but generally yields more easily interpretable output and much better sensitivity than BLAST1.

**BLAST FLAVORS**

BLAST programs come in five flavors:

- **blastp** compares an amino acid query sequence against a protein sequence database. Fully gapped alignments available in BLAST2.
- **blastn** compares a nucleotide query sequence against a nucleotide sequence database. Fully gapped alignments available in BLAST2.
- **blastx** compares a nucleotide query sequence translated in all reading frames against a protein sequence database. In BLAST2, has ‘in-frame’ gapped alignments and uses sum statistics to link alignments from different frames.
- **tblastn** compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames. In BLAST2, has ‘in-frame’ gapped alignments and uses sum statistics to link alignments from different frames.
- **tblastx** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that tblastx program cannot be used with the nr database on the BLAST Web page. Only ungapped alignment is available.

**BEAUTY HOMOLOGY SEARCH**

BLAST Enhanced Alignment Utility, an enhanced version of the NCBI's BLAST tool, searches three sequence databases and incorporates information on sequence family membership, the location of the conserved domains, and the locations of any annotated domains and sites directly into BLAST search results.
These enhancements make it much easier to detect weak, but functionally significant, matches in BLAST database searches. Results are shown in the Information Window. Save the results as a note for future reference.

**GenPept NR Blast Homology**

The GenPept NR service launches a BLAST homology search against a non-redundant protein database. GenPept, available through Entrez, is the standard protein flat file, the protein equivalent of GenBank.

**GenQuest Blast Homology**

GenQuest/Q server is an integrated interface to the sequence comparison server at the Oak Ridge National Lab designed for rapid and sensitive comparison of DNA and Protein sequence to existing DNA and Protein sequence databases and the rapid retrieval of the full database entries of any sequence found in the course of a search. The address is http://www.bis.med.jhmi.edu/Dan/gq/gq.form.html.

**SCOP Blast**

Find structural homologs for a protein sequence (via BLAST at SCOP, U.K.). The Structural Classification of Protein database, maintained at the Laboratory of Molecular Biology at Cambridge, UK, is a derivative of the Protein Database.

**Worldbase Records**

Sample records for BLAST searches are shown below. Be aware that the variable definitions – particularly the PATH= variable – are labile. Verify the address shown here when editing these records.

**Worldbase Record for Beauty**

```
# BEAUTY Search
SOURCE=beauty
INPUT=Seq
NAME=BEAUTY Homology Search
PATH=http://dot.imgen.bcm.tmc.edu:9331/cgi-bin/seq-search/protein-search.pl?seq_data=<%seq%>&seq_name=&email=&program_name=beauty_c_r_annot
SEQMAX=7000
```
Discovery Engine Reference Manual

SYNOPSIS=BLAST search with "BEAUTY" (BLAST Enhanced Alignment Utility) post-processing at Baylor. Provides line diagrams depicting hit alignments, as well as links for retrieving sequences.

TITLE=BEAUTY_Search_on_<<%s>>.note
VIEWER=document&

Worldbase RECORD FOR BLASTP
# Fix for new NCBI BLAST WWW page location.
SOURCE=blast-prot
INPUT=Seq
NAME=BLAST (Protein)

SYNOPSIS=Basic BLAST. Search for protein sequence homologs at NCBI by BLASTP against the non-redundant sequence database.
TITLE=Blast_Search_on_<<%s>>.note
VIEWER=document&

WHITESPACE=+

Worldbase RECORD FOR BLASTX HOMOLOGY ANNOTATIONS
# ANNOTATE DNA SEQUENCES BY HOMOLOGY TO PROTEIN SEQUENCES
SOURCE=ncbi-blastx
FILTER=a"$LOOK_DIR/exec/ncbi-blastx.awk"
INPUT=Seq
NAME=BLASTX Homology Annotations
OUTPUT=BLASTHomology
PATH=@http://www.ncbi.nlm.nih.gov:80/cgi-bin/BLAST/nph-blast??PROGRAM=blastx&DATALIB=nr&SEQUENCE=<<%s>>&EXPECT=default&CUTOFF=default&MATRIX=BLSUM62&STRAND=both&FILTER=default&DESCRIPTIONS
Discovery Engine Reference Manual

=10&ALIGNMENTS=100&ADVANCED=&PATH=
SEQMAX=15000
SERVICE=dna-annot-regular,dna-annot-full
STRIP=
SYNOPSIS=Find DNA sequence homologs by BLASTX on the non-redundant sequence database at NCBI.
TITLE=BLASTX Homology annotations for <<<%s>>>
**Discovery Engine Reference Manual**

SEQMAX=15000
SERVICE=strc-homology
STRIP=\?atgcu
SYNOPSIS=Find structural homologs for a protein sequence (via BLAST on the PDB).
TITLE=PDB BLAST Homologies for <<<%s>>>

**Worldbase RECORD FOR NR BLAST HOMOLOGY**

# BLASTP ANNOTATE PROTEIN SEQUENCES BY HOMOLOGY TO NR PROTEIN DB
# USE AS BACKUP IF FASTA FAILS US!
SOURCE=ncbi-blastp
FILTER=a"$LOOK_DIR/exec/ncbi-blastp.awk"
INPUT=Seq
NAME=NR BLAST Homology Annotations
OUTPUT=BLASTHomology
PATH=@http://www.ncbi.nlm.nih.gov:80/cgi-bin/BLAST/nph-blast??PROGRAM=blastp&DATALIB=nr&SEQUENCE=<<<%seq%s>>>&EXPECT=default&CUTOFF=default&MATRIX=BLOSUM62&STRAND=both&FILTER=default&DESCRIPTIONS=200&ALIGNMENTS=200&ADVANCED=&PATH=
SEQMAX=15000
SERVICE=nr-fasta&swiss-fasta&pir-fasta
STRIP=\?atgcu
SYNOPSIS=Find protein sequence homologs by BLAST against the non-redundant database at NCBI.
TITLE=NR BLAST Homologies for <<<%s>>>
OUTPUT=BLASTHomology
PATH=@http://scop.mrc-lmb.cam.ac.uk/scop/aln.cgi??loc=http://scop.mrc-
lmb.cam.ac.uk/scop/
&...
SEQMAX=1000
STRIP=}
SYNOPSIS=Search for structural homologs for a protein sequence using BLAST on the PDB via the GenQuest server at JohnsHopkins.
TITLE=PDB_Blast_Search_on_<<<%%s>>>.note
VIEWER=document

**BLOCKS HOMOLOGY SEARCH**

BLOCKS is a structure-motif based sequence-homology search available through the Baylor College of medicine. It compares the query sequence to the current database of protein blocks. Blocks are short, multiply-aligned ungapped, segments corresponding to the most highly conserved regions of proteins. Results will be seen in the Information Window.

**BLOCKS HOMOLOGY SEARCH**

# BLOCKS Search
SOURCE=blocks
INPUT=Seq
NAME=BLOCKS Homology Search
PATH=http://dot.imgen.bcm.tmc.edu:9331/cgi-bin/seq-search/protein-search.pl?seq_data=<<<%seq%$>>>&seq_name=&email=&program_name=blocks
SEQMAX=7000
STRIP=}
SYNOPSIS=Search for protein sequence homologs using the BLOCKS server at FHCRC via Baylor. This program can identify conserved motifs, even if they have no known function.
TITLE=BLOCKS_Search_on_<<<%%s>>>.note
CLUSTAL SEQUENCE ALIGNMENT

This server provides access to CLUSTAL for multiple sequence alignments. This can provide a different alignment, which may guide you in optimizing your alignment in GeneMine.


Worldbase RECORD

FOR CLUSTAL

# CLUSTAL Alignment
SOURCE=clustal
INPUT=Seq
NAME=CLUSTAL Sequence Alignment
PATH=http://dot.imgen.bcm.tmc.edu:9331/cgi-bin/multi-align/clustalw.pl??seq_data=<<<>F1;%%s
%title%s
%seq%s
*
>>>&program_name=clustalw
SEQMAX=20000
STRIP=]
SYNOPSIS=Send sequences (you need to select at least two) to be aligned by Clustal 1.6 through the Baylor College of Medicine site -- http://dot.imgen.bcm.tmc.edu:9331/. Results are returned in the Information window, and may be saved as a hypernote.
TITLE=Clustal Alignment of <<<%%s>>>
VIEWER=document

dbEST INFORMATION

dbEST (Nature Genetics 4:332-3;1993) is a division of GenBank that contains sequence data and other information on "single-pass" cDNA sequences, or Expressed Sequence Tags, from a number of organisms.
**DBSTS ELECTRONIC PCR**

dbSTS is an NCBI resource that contains sequence and mapping data on short genomic landmark sequences or Sequence Tagged Sites (Olson et al., 1989). Annotation in dbSTS is comprehensive and includes detailed contact information about the contributors, experimental conditions and genetic map locations. In addition, NCBI periodically updates putative homology assignments using the BLAST family of programs (Altschul et al., 1994).

PCR-based sequence tagged sites (STTs) have been used as landmarks for construction of various types of genomic maps. Using "electronic PCR" (e-PCR), these sites can be detected in DNA sequences, potentially allowing their map locations to be determined.

**Worldbase RECORDS**

**Worldbase RECORD FOR DBEST TRANSCRIPTION**

```
SOURCE=ncbi-est-blastn
FILTER=a"$LOOK_DIR/exec/ncbi-blastn.awk"
INPUT=Seq
NAME=dbEST Transcription Annotations
OUTPUT=BLASTHomology
PATH=@http://www.ncbi.nlm.nih.gov:80/cgi-bin/BLAST/nph-blast??PROGRAM=blastn&DATALIB=dbest&SEQUENCE=<<<<%seq%s>>>>&EXPECT=default&CUTOFF=default&MATRIX=BLOSUM62&STRAND=both&FILTER=default&DESCRIPTION=10&ALIGNMENTS=100&ADVANCED=&PATH=
SEQMAX=30000
SERVICE=dna-annot-regular,dna-annot-full
STRIP=]
SYNOPSIS=Find expression information for a DNA sequence by BLASTN on the EST database at NCBI.
TITLE=dbEST Transcription annotations for <<<%%s>>>
```

**Worldbase RECORD FOR DBEST EXPRESSION**

```
# ANNOTATE PROTEIN SEQUENCES BY EXPRESSION IN dbest
SOURCE=ncbi-est-tblastn
```
FILTER=a"$LOOK_DIR/exec/ncbi-tblastn.awk"
INPUT=Seq
NAME=dbEST Expression Annotations
OUTPUT=BLASTHomology
PATH=@http://www.ncbi.nlm.nih.gov:80/cgi-bin/BLAST/nph-blast??PROGRAM=tblastn&DATALIB=dbest&SEQUENCE<<<%seq%s>>>&EXPECT=default&CUTOFF=default&MATRIX=BLOSUM62&STRAND=both&FILTER=default&DESCRIPTIONS=10&ALIGNMENTS=100&ADVANCED=&PATH=
SEQMAX=30000
SERVICE=prot-annot-regular,prot-annot-full
STRIP=]?
atgcu
SYNOPSIS=Find expression information for a protein sequence by TBLASTN on the EST database at NCBI (dynamically translated in all 6 frames).
TITLE=dbEST Expression annotations for <<<%%s>>>
**Worldbase RECORD FOR DBEST PROTEIN CLUSTERING**

# ANNOTATE PROTEIN SEQUENCES BY EXPRESSION IN dbest, STRINGENT

SOURCE=ncbi-est-tblastn-hs
FILTER=a"$LOOK_DIR/exec/ncbi-tblastn-stringent.awk"
INPUT=Seq
NAME=dbEST Clustering, protein level
OUTPUT=BLASTHomology
PATH=@http://www.ncbi.nlm.nih.gov:80/cgi-bin/BLAST/nph-blast??PROGRAM=tblastn&DATALIB=dbest&SEQUENCE=<<<%seq%s>>>&EXPECT=default&CUTOFF=default&MATRIX=BLOSUM62&STRAND=both&FILTER=default&DESCRIPTIONS=10&ALIGNMENTS=100&ADVANCED=&PATH=SEQMAX=30000
STRIP=\?atgcu
SYNOPSIS=Find expression information for a protein sequence by TBLASTN on the EST database at NCBI, with stringent filtering of results.
TITLE=dbEST Expression annotations for <<<%s>>>

**Worldbase RECORD FOR DBSTS**

# ANNOTATE DNA SEQUENCES BY STS MAP LOCATION

SOURCE=ncbi-sts-epcr
FILTER=a"$LOOK_DIR/exec/ncbi-epcr.awk"
INPUT=Seq
NAME=dbSTS Electronic PCR
OUTPUT=SeqDBAnnotation
PATH=@http://www.ncbi.nlm.nih.gov:80/cgi-bin/STS/nph-sts/??INPUT_TYPE=Sequence+in+FASTA+format&SEQUENCE=<<<%seq%s>>>&ORGANISM=All+Organisms&DETACHED=on
SEQMAX=30000
SERVICE=dna-annot-full
STRIP=\?
SYNOPSIS=Search for DNA sequence chromosomal map locations by searching against the STS database at NCBI.
TITLE=dbSTS Electronic PCR Annotations for <<<%s>>>
FASTA HOMOLOGY

For a review of the FASTA sequence similarity calculation and scoring method, see http://www-biology.ucsd.edu/others/dsmith/fast.html.

NR means non-redundant.

Worldbase RECORDS

Worldbase Record for FASTA NR Homology EERIE
# FASTA Protein Homology Searches: Total NR Sequence Set
SOURCE=eerie-nr-fasta
FILTER=a"$LOOK_DIR/exec/eerie-nr.awk"
HISTORY=nr-fasta
INPUT=Seq
NAME=FASTA NR Homology Annotations EERIE
OUTPUT=BLASTHomology
PATH=@http://vega.crbm.cnrs-mop.fr/bin/nph-fasta_query.pl??br=no_script&ou=www&ad=&db=nr+N&pr=fasta&ml=&kt=1&ma=BLAST62&sc=200&al=200&co=seq,&qu=<<<%seq%s>>>&ty=Protein
STRIP]=?atgcu
SYNOPSIS=Find protein sequence homologs (via FASTA on the non-redundant database at GeneStream [http://vega.crbm.cnrs-mop.fr]).
TITLE=FASTA Homology Annotations for <<<%s>>>

Worldbase Record for FASTA Swissprot Homology EERIE
# SWISSPROT Sequence Set
SOURCE=eerie-swiss-fasta
FILTER=a"$LOOK_DIR/exec/eerie-fasta.awk"
HISTORY=swiss-fasta
INPUT=Seq
NAME=FASTA Swissprot Homology Annotations EERIE
OUTPUT=BLASTHomology
PATH=@http://vega.crbm.cnrs-mop.fr/bin/nph-fasta_query.pl??br=no_script&ou=www&ad=&db=swissprot+S&pr=fasta&ml=&kt
Discovery Engine Reference Manual

=1&ma=BLOSUM62&sc=200&al=200&co=seq,&qu=<<<%seq%s>>>&ty=Protein
SERVICE=swiss-fasta&
STRIP=!?atgcu
SYNOPSIS=Find protein sequence homologs (via FASTA on the SwissProt database at GeneStream [http://vega.crbm.cnrs-mop.fr]).
TITLE=FASTA Homology Annotations for <<<%s>>>  

Worldbase RECORD FOR GQ FASTA HOMOLOGY
# USE GENQUEST AS BACKUP SERVICE FOR SWISS PROT
SOURCE=gq-swiss-annot-auto
FILTER=a"$LOOK_DIR/exec/genq-fasta.awk"
INPUT=Seq
NAME=GQ FASTA Homology Annotations
OUTPUT=BLASTHomology
PATH=@@http://www.gdb.org/bio/Quest/gq-cgi.pl??TYPE=Protein&TARGET=SwissProt&METHOD=FASTA&MATRIX=Blosum&Blosu
m-param=62&PAM-param=130&Score=40&Align=40&gap=13&Align-
type=Local&Filter=No&Frame=1&Sequence=<<<%seq%.999s>>>  
SEQMAX=1000
SERVICE=swiss-fasta&
STRIP=!?atgcu
SYNOPSIS=Find protein sequence homologs (via FASTA on the SwissProt database at GenQuest [http://www.gdb.org/Dan/gq/gq.form.html]).
TITLE=FASTA Homology Annotations for <<<%s>>>  

FEATURE INFORMATION

CLONING VECTORS

Access to Cloning Vectors, a database containing information on plasmids, phages, etc. used for cloning. When searching for plasmids, add a "p" to the beginning of your search query (e.g., "pbluescript" rather than "bluescript").
Worldbase RECORD

FOR CLONING VECTORS

# Queens Univ. Alternate Vector DB
SOURCE=quVect
AND=;
NAME=Cloning Vectors (Alternate)
OR=,
PATH=http://biol.gis.queensu.ca/cgi-bin/aglimpse/02?query=%key%&case=on&whole=on&errors=0&maxfiles=10&maxlines=3
SYNOPSIS=Search for Cloning Vectors at the VectorDB. Supports "and" to limit searches (ie. "lambda and YES").
TITLE=Search Vector DB for %key%
VIEWER=document
WHITESPACE=+

DNA REPEAT DETECTION

RepeatMasker, accessed through the Baylor College of Medicine, which references the Genome center at the University of Washington (http://ftp.ge
nome.washington.edu), screens DNA sequences in FASTA format against a library of repetitive elements and returns a masked query sequence ready for da
tabase searches as well as a table annotating the masked regions.


For information on cutoffs and how to read the results, see the documentation page at http://ftp.genome.washington.edu/RM/RM_details.html.

Worldbase RECORD

FOR DNA REPEAT DETECTION

SOURCE=bcm-dna-repeat
FILTER=a"$LOOK_DIR/exec/repeat_masker.awk"
INPUT=Seq
NAME=DNA Repeat Detection
Discovery Engine Reference Manual

OUTPUT=SeqDBAnnotation
PATH=@http://dot.imgen.bcm.tmc.edu:9331/cgi-bin/seq-util/seq-util.pl??seq_data=<<<%seq%s>>>&program_name=repeat_masker
SEQMAX=30000
SERVICE=dna-annot-full
STRIP=]
SYNOPSIS=Sends a DNA sequence for detection of interspersed repeats and regions of low complexity (see http://ftp.genome.washington.edu/RM/RM_details.html).
TITLE=DNA Repeat Masking of <<<%%s>>>

DOMAINS


Results are returned as automatic annotations following sequence queries with [which???] Workup levels.

Worldbase RECORD

FOR PRODOM ANNOTATIONS

SOURCE=prodom-swiss-annot-auto
INPUT=Seq
NAME=Prodom Annotations
OUTPUT=ProdomAnnotation
PATH=@http://www.embl-heidelberg.de/srs/srsc?[swissdom-id=<<<%s>>>
SERVICE=swiss-annot-auto
SYNOPSIS=Find known sequence features for a protein sequence from PRODOM (via SRS at EMBL).
TITLE=Domain Search on <<<%s>>>
**EUKARYOTIC PROMOTERS**

Search for promoter sequences of Eukaryotic genes, through GDB, which references http://benpc.bionet.nsc.ru/TRRD/.

Tips for searching: Try searching for words instead of acronyms. For example, hsp has been translated to Heatshock Protein, so search for "heatshock". You may use booleans ("and", "or", "not"), wildcards ("*") and phrase searches (" ").

**MHC PEPTIDES**

Access to MHCPEP, a curated database at The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia. The database comprises over 13423 peptide sequence entries (indexed as of 15-Oct-1997) known to bind MHC molecules. Entries are compiled from published reports as well as from direct submission of experimental data. Each entry contains the source protein (when known), an estimate of binding affinity and critical anchor residues (if identified), and is fully referenced. The present format of the database allows text string matching searches.

Supports Booleans and wildcards (i.e., "recept* and alpha").

**Worldbase RECORD FOR MHC PEPTIDES**

# MHC peptides DB Text Search
SOURCE=mhcpep
NAME=MHC Peptides
PATH=gopher://wehil.wehi.edu.au/77/MHCPEP.DB/.index/index?%key%
SYNOPSIS=Search for MHC peptides by keyword in the database at WEHI in Australia. Supports Booleans and wildcards (i.e. "recept* and alpha").
WILDCARD=* 

**OPEN READING FRAMES**

Search for Open Reading Frames (ORFs) in two different modes:

- Beginning with any amino acid, and at least 25 amino acids long.
Discovery Engine Reference Manual

- Beginning with methionine, and at least 100 amino acids long.

The server is at ALCES at the University of Minnesota (http://alces.med.umn.edu/webtrans.html).

A third ORF finder is available through NCBI. The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server.

Results are returned as annotations following Light, Regular or full workups.

Worldbase RECORDS

Worldbase RECORD FOR ORFS (ANY, >25AA)
# Find ORFs. Starts with Any aa, greater than 25 amino acids.
SOURCE=umORFA
INPUT=Seq
NAME=Find ORFs (Any, >25aa)
PATH=http://alces.med.umn.edu/bin/webcuse??imode=Raw&name=&rawseq=<<<%seq$s>>>&left=0&right=0&size=25&co
deflag=Standard&mflag=Any&uflag=All&omode=Translate
SEQMAX=30000
SYNOPSIS=Search for Open Reading Frames (ORFs) beginning with any amino acid, and at least 25 amino acids long, in a DNA sequence using the server at ALCES at the University of Minnesota (http://alces.med.umn.edu/webtrans.html).
TITLE=ORF (Any) Search on <<<%%s>>>
VIEWER=document

Worldbase RECORD FOR ORFS (MET, >100AA)
# Find ORFs. Starts with Met, greater than 100 amino acids.
SOURCE=umORFM
INPUT=Seq
NAME=Find ORFs (Met, >100aa)
Discovery Engine Reference Manual

PATH=http://alces.med.umn.edu/bin/webcuse?imode=Raw&name=&rawseq=<<<%seq%s>>>&left=0&right=0&size=100&codeflag=Standard&mflag=Met&uflag=All&omode=Translate
SEQMAX=30000
SYNOPSIS=Search for Open Reading Frames (ORFs) beginning with methionine, and at least 100 amino acids long, in a DNA sequence using the server at ALCES at the University of Minnesota (http://alces.med.umn.edu/webtrans.html).
TITLE=ORF (Met) Search on <<<%%s>>>
VIEWER=document

Worldbase RECORD FOR NCBI ORF FINDER
SOURCE=ncbi-gorf
FILTER=p"$LOOK_DIR/exec/ncbi-gorf.pl"
INPUT=Seq
NAME=NCBI ORF Finder
OUTPUT=SeqDBAnnotation
PATH=@http://www.ncbi.nlm.nih.gov/cgi-bin/gorf/orfig??acc=&seq=<<<%seq%s>>>&from=&to=&%5C%22gcode%5C%22=1+Standard
SEQMAX=30000
SERVICE=dna-annot-light, dna-annot-regular, dna-annot-full
STRIP=]
SYNOPSIS=Find predicted Open Reading Frames for a DNA sequence via the ORF finder at NCBI.
TITLE=NCBI ORF Finder Annotations for <<<%%s>>>

PCR PRIMERS

Predicted PCR primers for long genomic sequences are generated by primertx, a variant of the xprimer program, and are retrieved as automatic annotations when Regular or Full Workups are specified. The service is provided by the Baylor College of Medicine, which references the ALCES site in Minnesota.
Worldbase RECORD

Worldbase RECORD FOR PCR PRIMERS

SOURCE=bcm-pcr-primer
FILTER=a"$LOOK_DIR/exec/primertx.awk"
INPUT=Seq
NAME=PCR Primer Generation
OUTPUT=SeqDBAnnotation
PATH=@http://dot.imgen.bcm.tmc.edu:9331/cgi-bin/seq-util/seq-util.pl??seq_data=<<<%seq%s>>>&program_name=primertx
SEQMAX=30000
SERVICE=dna-annot-regular,dna-annot-full
STRIP=?
SYNOPSIS=Find predicted PCR primers for a DNA sequence at Alces via Baylor --http://dot.imgen.bcm.tmc.edu:9331/seq-util.
TITLE=PCR Primers for <<<%s>>>
Worldbase RECORDS

RECORD FOR PROSITE TEXT RETRIEVAL AT ExPASY

This record launches a keyword search of the ExPASY database at Geneva and returns results as text to the Information window. Select "Prosite motif" in the File:New:Keyword Search... drop-down list. Use the context-sensitive menu to follow links relevant to the search results.

# PROSITE MOTIFS TEXT SEARCH
SOURCE=prosite-expasy-key
NAME=Prosite Motifs
PATH=http://expasy.hcuge.ch/cgi-bin/prosite-search-ful?%key%
SYNOPSIS=Search for a protein motif by name or keyword (ie. "zinc and finger").
TITLE=Prosite Search on %key%

FOR PROSITE AUTO ANNOTATIONS AT EBI

This record launches an automatic sequence search of the EBI database at the Sanger Center and returns results as an annotation appended to the sequence in the Sequence window.

SOURCE=ebi-prosite-annot
FILTER=a"$LOOK_DIR/exec/ebi-prosite.awk"
HISTORY=prosite-annot
INPUT=Seq
NAME=EBI Prosite Motif Detection
OUTPUT=PrositeMotif
PATH=@http://www.ebi.ac.uk/htbin/prosite_input_parser??title=&full=no&seq=<<<<%seq%s>>>
SEQMAX=30000
SERVICE=prosite-annot&
STRIP=]atgcu
SYNOPSIS=Find motifs in a protein sequence (via Prosite Server at EBI).
TITLE=Prosite Scan of <<<%s>>>

181
**Discovery Engine Reference Manual**

**FOR PROSITE AUTO ANNOTATIONS AT ExPASy**

This record launches the same type of automatic sequence search as above, of the ExPASy database, and is used by Full, Regular or Light workups.

SOURCE=expasy-prosite-annot  
HISTORY=prosite-annot  
INPUT=Seq  
NAME=Prosite Motif Detection  
OUTPUT=PrositeMotif  
PATH=@http://expasy.hcuge.ch/cgi-bin/scanprosite?1??ID=&SEQ=<<<%seq%s>>>&box=ok  
SEQMAX=30000  
SERVICE=prot-annot-light,prot-annot-regular,prot-annot-full  
STRIP=\?[atgcu  
SYNOPSIS=Find known motifs in a protein sequence (via ExPaSy Prosite Server).  
TITLE=Prosite Scan of <<<%%s>>>  

**RESTRICTION SITES**

Find restriction sites in a DNA sequence using the Webcutter program provided by the Baylor College of Medicine. A mirror site is at the University of Goteborg in Sweden. Results are returned as automatic annotations in Regular and Full Workup sequence queries.

**Worldbase RECORDS**

**FOR WEBCUTTER UNIQUE RESTRICTION SITES**

SOURCE=bcm-webcutter  
FILTER=a"$LOOK_DIR/exec/webcutter.awk"  
INPUT=Seq  
NAME=All Restriction Sites  
OUTPUT=SeqDBAnnotation  
PATH=@http://dot.imgen bcm tmc edu:9331/cgi-bin/seq-util/seq-
util.pl??seq_data=<<<%seq%s>>>&program_name=webcutter
SEQMAX=30000
STRIP=][
SYNOPSIS=Find all restriction sites in a DNA sequence using WebCutter via the Baylor site -- http://dot.imgen.bcm.tmc.edu:9331/seq-util/seq-util.html.
TITLE=Webcutter Restriction Analysis of <<<%%s>>>

FOR WEBCUTTER (BACKUP)
SOURCE=medkem-webcutter
FILTER=a"$LOOK_DIR/exec/webcutter.awk"
INPUT=Seq
NAME=Webcutter Unique Restriction Sites
OUTPUT=SeqDBAnnotation
PATH=@http://www.medkem.gu.se/cgi-bin/cutter/cutter??title=Untitled+sequence&sequence=<<<%seq%s>>>&number=all&entered=&rangelow=&rangehigh=&table=&number=unique&entered=&stranegold=`&rangehigh=&alphapos=alph&enzymes=longenz&sitelength=6
SEQMAX=30000
SERVICE=dna-annot-full
STRIP=][
SYNOPSIS=Find unique restriction sites in a DNA sequence using WebCutter via the Baylor site -- http://dot.imgen.bcm.tmc.edu:9331/seq-util/seq-util.html.
TITLE=Webcutter Restriction Analysis of <<<%%s>>>

SECONDARY STRUCTURE

The Secondary Structure Prediction Program (SSP) predicts α-helix and β-strand segments of globular proteins. The server is at the Department of Cell Biology, Baylor College of Medicine (BCM).

Results are returned as automatic annotations for sequence queries for Light, Regular and Full Workups.
Worldbase RECORDS

FOR PSSP REPORT SERVER

SOURCE=pssp-report
FILTER=s"sed 's/<.*>//g'"
INPUT=Seq
NAME=PSSP Report Server
PATH=@http://dot.imgen.bcm.tmc.edu:9331/cgi-bin/seq-search/struc-predict.pl?seq_data=<<<%seq%s>>>&seq_name=<<<%%s>>>&email=&program_name=pssp_ssp
SEQMAX=7000
SERVICE=secstrc-note
STRIP=X)?atgcu
TITLE=PSSP Predicted Secondary Structure for <<<%%s>>>
UI=hide

FOR PSSP PREDICTION SERVER

SOURCE=pssp-seq-annot-auto
HELP=PATH=http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html
INPUT=Seq
NAME=PSSP Prediction Server
OUTPUT=SecStrcAnnotation
PATH=@http://dot.imgen.bcm.tmc.edu:9331/cgi-bin/seq-search/struc-predict.pl?seq_data=<<<%seq%s>>>&seq_name=noname&email=&program_name=pssp_ssp
SEQMAX=2000
SERVICE=prot-annot-light,prot-annot-regular,prot-annot-full
STRIP=X)?atgcu
SYNOPSIS=Protein Secondary Structure Prediction server, located at the Baylor College of Medicine in Houston.
TITLE=PSSP Predicted Secondary Structure for <<<%%s>>>


TRANSCRIPTION FACTORS

Search for transcription factors and find their DNA binding sequence using the Transcription Factors database. Searches for SEQUENCE matches to this database can be done through gopher by using either:

- The exact, entire sequence pattern as it appears in this database (i.e., no expansion of IUPAC ambiguous codes), for example "KGGCGGRRY"
- The partial word search option of this Gopher server (.....the * wildcard can only be used to expand the "word" to the right), for example "KGGCGG*"
- The boolean operators "and", "or", and "not" are recognized as operators. If no specific separator is used between multiple words the "or" operator is assumed. The Gopher protocol does not provide the ability to search for specific DNA or Protein sequence patterns within a dataset. Your search results will be displayed in the Network Queries dialog window. You can retrieve any of the results which interest you by selecting it and clicking Retrieve. The Transcription Factor information will be displayed in the Information Window.

Worldbase RECORD

FOR TRANSCRIPTION FACTORS

# Transcription Factor DB Text Search
SOURCE=tfd
NAME=Transcription Factors
PATH=gopher://gopher.nih.gov:70/77/gopherlib/indices/tfd/index?%key%
SYNOPSIS=Find transcription factor binding sites and related information by keyword search of the database at NIH. Supports Booleans and wildcards (i.e. "jun and nuc*").
WILDCARD=* 

TRANSLATE IN 6 FRAMES

This valuable utility searches raw stretches of DNA for open reading frames. Using the n2a (Nucleotide to Amino Acid Conversion Utility) program, the Translate in 6 Frames tool examines your DNA sequence in all three reading frames, as well as the complementary strand in all three reading frames. The server returns a copy of the DNA sequence you sent, followed by six translated amino
acid sequences. The amino acid sequences are represented by single-letter amino acid codes with two exceptions, and stop codons are shown with an asterisk ".

These results are displayed in the Information Window. This tool is useful when searching for reading frames in large stretches of DNA. To read your sequence into the Sequence Window, choose New Sequence from the sequence menu, and copy the desired sequence from the Information Window into the dialog box, or use "Import Alignment" for DNA sequences already aligned in a known format.

**Worldbase RECORD**

**FOR TRANSLATE IN 6 FRAMES**

SOURCE=trxlt
INPUT=Seq
NAME=Translate in 6 Frames
PATH=http://www.ibc.wustl.edu/n2a.cgi?n=<<<%seq%s>>>
SEQMAX=2000
SYNOPSIS=Send a DNA sequence to be translated in all 6 frames by the server at Washington University, St. Louis (WUSTL). Stop codons are shown as asterisks ("*"). The information is returned in the Information window, and may be saved as a hypernote.
TITLE=Six Frame Translation of <<<%s>>>
PCR and sequenc* not DNA

will find messages that contain the text "PCR" and any words starting with "sequenc" (sequence, sequencing, etc.) but exclude messages containing the text "DNA." Terms do not occur in any specific order in the message.

Worldbase RECORD

FOR BIOTECHNIQUES

# BioTechniques-BioNet News Group Search
SOURCE=biotech-arch
NAME=BioTechniques
SYNOPSIS=Search by keyword for Biotechniques articles. Combine words with "and" to limit the number of results (i.e. "electro* and gel").
TITLE=BioTechniques Search on %key%
WILDCARD=*  

CHEMFINDER DATABASE

Contents of the ChemFinder WebServer, the WWW version of Chemical abstracts Service, provided by CambridgeSoft. The chemFinder database contains information about chemical compounds, including physical property data and 2D chemical structures. In Discovery Engine search by the name of the compound or by its CAS number. To search by formula, molecular weight, boiling point and structure, access the ChemFinder site directly from Netscape, then use the "Open from Netscape" command to bring the results into Discovery Engine.

Worldbase RECORD

FOR CHEMFINDER

# ChemFinder (CambridgeSoft)
SOURCE=chemdb
NAME=ChemFinder Database
PATH=http://chemfinder.camsoft.com/cgi-win/
**Discovery Engine Reference Manual**

cfserver.exe??nametype=contains&name=%key%
SYNOPSIS=Search by keyword for information on chemicals using CambridgeSoft's ChemFinder -- i.e. "*benzene"
TITLE=ChemFinder Search on %key%
VIEWER=document
WHITESPACE=+
WILDCARD=* 

**EC Enzyme Database**

The EC ENZYME database, accessed through ExPASy (http://expasy.hcuge.ch) is a repository of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided.

Data types include:
- EC number
- Recommended name
- Alternative names (if any)
- Catalytic activity
- Cofactors (if any)
- Pointers to the SWISS-PROT entry(s) that correspond to the enzyme (if any)
- Pointers to disease(s) associated with a deficiency of the enzyme (if any)

Search using booleans ("and", "or", "not"), nested booleans, wildcards ("*"), and phrase searches (" "). The maximum number of hits is 200.

**Worldbase Record**

**For EC Enzyme Database**

# ENZYME FAMILIES TEXT SEARCH
SOURCE=ecenzyme-expasy-key
NAME=EC Enzyme Database
PATH=http://expasy.hcuge.ch/cgi-bin/enzyme-search-de?%key%
SYNOPSIS=Search for an enzyme by keyword (ie. "kinase"). No support for Booleans or wildcards.
TITLE=EC Enzyme Search on %key%

**J. BIOL. CHEM FULL TEXT**

Provides online access to the most recent 14 months contents of the J. of Biological Chemistry, a weekly publication of The American Society for Biochemistry and Molecular Biology.

Boolean searches are supported. Only the best 30 matches are shown.

**Worldbase RECORD**

**JBC ONLINE (LAST 14 MONTHS)**

# JBC search 1995-1997 (no more 'last 14 months' option).
SOURCE=JBCfull
NAME=JBC Online (last 14 months)
PATH=http://www.jbc.org/cgi/search?sendit=Start+Search&author1=&author2=&titleabstract=&fulltext=%key%&fyear=1995&tyear=1997&hits=100
SYNOPSIS=Search for Journal of Biological Chemistry articles by keyword. Supports Booleans and wildcards (i.e. "homeo* and xenopus"). Maximum 100 hits returned.
TITLE=JBC Search on %key%
WHITESPACE=+
WILDCARD=* 

**MEDLINE REFERENCES**

MEDLINE (MEDlars onLINE) is the National Library of Medicine's (NLM) premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. The MEDLINE file contains bibliographic citations and author abstracts from Each MEDLINE record is identified with a unique identifying number called a MED-
LINE UID (MUID in PubMed). MEDLINE records are incorporated into PubMed weekly, and are also assigned a PubMed unique identifier (PMID).

The medical citations available through WWW Entrez include the entirety of the MEDLINE database (over 3,800 current biomedical journals published in the United States and 70 foreign countries), plus citations from journals not indexed by MEDLINE. This set contains approximately 8.7 million citations. In addition, the full text of many journals' articles are available through PubMed by means of links to the publishers of those journals.

Many of the PubMed citations in WWW Entrez contain protein or nucleotide sequence information; these citations are linked to their constituent sequences via Entrez, permitting you to move easily from a view of the PubMed abstract to one of the sequence(s) it contains.

In addition, each PubMed citation has been compared to many of the other PubMed citations in the Entrez database using an algorithm that evaluates the similarity of the text and MeSH terms found in those citations. The most similar citations, called "neighbors", can then be viewed through Entrez. This facility permits you to find a large number of citations that fall into your area of interest once you have found a few relevant citations, and increases the power of your searching dramatically.

Worldbase RECORD

FOR MEDLINE

# Entrez MEDLINE
SOURCE=entrez-medline
AND=[text]+--AND--+
NAME=MEDLINE References
NOT=[text]+--BUT+NOT--+
OR=[text]+--OR--+
SYNOPSIS=Search for literature by keywords through Entrez MEDLINE. Supports Booleans and wildcards (i.e. "hiv and integra").
TITLE=MEDLINE Search on %key%
WHITESPACE=+
WILDCARD=*
NIH GRANTS

Access to the NIH CRISP (Computer Retrieval of Information on Scientific Projects) system, a major biomedical database containing information on research ventures supported by the United States Public Health Service (US-PHS). Helpful hints when searching CRISP on Gopher:

- When searching by project number, do NOT enter any spaces, for example: (R01DK36081)
- To search for a particular subproject within a program project, use the boolean operator "and", for example: (P01DC00036 and 0005)
- Use "and" when searching for an investigator by first and last name, for example: (Smith and John)

They come from gopher://gopher.nih.gov, but this site doesn’t work.

Worldbase RECORD

FOR NIH GRANTS

SOURCE=nih-crisp
NAME=NIH Grants
PATH=gopher://gopher.nih.gov:70/77/gopherlib/indices/crisp/index?%key%
SYNOPSIS=Search for NIH grants through the NIH gopher CRISP service. Supports Booleans and wildcards (i.e. "angio* and retinal").
WILDCARD=*  

SEQUENCE ANALYSIS REFS

As a keyword search, launches query to the ExPASy server at Geneva and returns the result as text to the Information window.

[971016SeqAnalRef]

As part of Full Workup sequence query, returns the result to the Sequence window adjacent to the sequence. View SeqAnalRefs in list format by clicking the [lit] icon.
**Discovery Engine Reference Manual**

This database contains bibliographic references pertinent to sequence analysis, including mathematical and computer analysis of biomolecular sequences. Most entries belong to one of the following categories:

- Algorithms for protein and nucleic acid sequence analysis: primary, secondary and tertiary structure analysis; pattern matching; similarity searches; alignments, etc.
- Algorithms for sequence-based phylogenetic analysis.
- Description of biopolymer data banks: nucleic acid, protein, tertiary structure, carbohydrates, etc.
- Description of software packages.
- Description of on-line services for molecular biologists.

Use a single keyword (i.e., "anneal" finds anneal, annealed, annealing, ...). Does not support Booleans (no AND or NOT, etc.).

**Worldbase RECORD**

**FOR SEQUENCE ANALYSIS REFS**

```
# Sequence Analysis Refs
SOURCE=seqanalref
NAME=Sequence Analysis Methods
PATH=http://expasy.hcuge.ch/cgi-bin/seqanalr-search-ful?%key%
SYNOPSIS=Search SeqAnalRef at ExPASy by keyword for citations on sequence analysis methods and algorithms. Use a single keyword (i.e. "anneal" finds anneal, annealed, annealing, ...). Does not support Booleans (no AND or NOT, etc).
TITLE=Seqanal Search on %key%
WILDCARD=* 
```

**U.S. PATENTS**

This server provides online access to the Commerce Department's Patent and Trademark Office 'Patent Database'. Search the Patent Bibliographic data from 1976 to the present. The Patent Office only supplies patent abstracts only through this server. Supports Booleans and wildcards (i.e., "dna and immuno*"). You
may also use field identifiers such as Assignee name (i.e., "an/ibm"). Only the top 50 hits are shown.

Worldbase RECORD

FOR U.S. PATENTS

# Update for change to the US Patents DB site setup.
SOURCE=patents
NAME=U.S. Patents
NOT=+ANDNOT+
PATH=http://patents.cnidr.org/cgi-bin/adv_srch3?/pto7/NEW/
INDEX??DBSELECT2=SPECIFY&DBSELECT=76-97&ADVANCED=%key%&RANKTYPE=CHRON&ELEMENT_SET=FT
SYNOPSIS=Search for patents by keyword at the USPTO (CNIDR) from 1976-1997. Supports Booleans and wildcards (i.e. "dna and immuno*"). You may also use field identifiers such as Assignee name (i.e. "an/ibm"). Only the top 50 hits are shown.
TITLE=U.S. Patent Search on %key%
WHITESPACE=+
WILDCARD=*

Discovery Engine Reference Manual
Index

Symbols
/local directory 76
| symbol 8, 134

Numerics
10-25% rule 115
50% identity rule 111
50% rule 6, 9, 111

A
About this Annotation 8, 11, 18, 31, 33, 49, 51
About this annotation 10, 116, 144
about-to-be annotated DNA sequence 54
about-to-be annotated protein sequence 39
Active site 136
active site 7
Active site annotations 42
Add Annotation 20, 30
additive filtering 66, 68
alignment 38, 39, 57
alignment display 28
And, Or, Not fields 82
Annotated DNA sequence 55
annotation 30
   calculated 5
   categories of 6
   display 26
   display by category 28
   display individual 28, 29
   formats 18
   icons 28, 41
   link 31
   not available 41
   pop-up menu 32, 33
   record 56, 57
   reference 52
   style 32
   uncertain a. italicized 5
   user-created 20
   annotation classes 41
   not displayed 27
   Annotation display 110
   Annotation fields 103
   annotation link 20, 30
   list 31
   not available for core homology 15, 123
   Annotation pop-up menu 33
   Annotation Styles 32
   Annotation Toolbar 27, 40
   literature icon not included 18, 144
   Annotations
   validity of method 47
   Application field 86
   Application recognition 71
   APPLICATION= 76
   Applications
   pop-up menu 85
   automated annotations 3, 4, 35
   average sequence identity 119
   awk script 84

B
back arrow 37
Backup servers 69
backup SERVICE
   & code 79
Baylor College of Medicine 16, 140
BEAUTY homology 162
BioTechniques 187
BLAST 4, 5, 9, 11, 12, 13, 41, 110, 116, 119, 120
   Entrez, SRS links 165
   flavors 162
   search 161
   BLAST 2 96
Index

BLAST alignment scoring 96
BLAST/FASTA 12, 120
BLAST/FASTA output 93
BLAST2 162
BLASTHomology
   Header formats 104
BLASTHomology Hits formats 108
BLASTX-MODE field 105
BLOCKS Homology Search 168
blue link 12, 120
BRCA1 52
   ORF annotation 57
buffer size 70

cut-off 94

data mining 51
data types
   recognition of 85
DATABASE field 105
dbEST 13, 94, 120, 121
   DNA clustering 171
   expression 170
   information 169
   protein clustering 172
   transcription 170
dbSTS 172
   electronic PCR 170
Define new source 89
disease association 14, 19, 139
diseases.hot 19
diseases.hot dictionary 139
display attributes 49
disulfide bridge annotation 17
disulfide bridge annotations 46
DNA
   repeat detection 175
to protein annotation 55
domain annotation 7, 42, 136

cache hash key 71
cached hypernote 116
caching 66, 70
CARA 80
ChemFinder Database 187
Choose Sequence(s) to Show 38, 53, 54
client-side processing 4, 41, 93
Cloning Vectors database 174
close homologs 12, 24
Close Homology 93
Closely related homolog annotations 45
Closely Similar Homology 122
CLUSTAL Sequence Alignment 169
cluster
   family 6
cluster leader 117
clustering
   minimum percent identity 12
Constraints on queries 64
carbohydrate-sensitive menu 38, 53
core homology 9, 10, 15, 111, 122
correspondence between annotation classes 47
cross-check annotations 47
customized information retrieval 147
Customizing Worldbase 89

196
Index

EST
  data, clustering schemes for 13
  file header 13
  homology annotations 46
EST cluster 122
EST data
  clustering schemes for 121
EST Expression 6, 13, 93, 120
  dictionary 13
EST file header 120
Eukaryotic Promoters 177
EVENT= 76, 88
ExPASy 4, 14, 138
EXPECT field 108
expect value
  same as p-value 104
expression annotation 19
expression.hot 19
EXPRESSION-MODE field 106

F
false positive 11, 133
family
  cluster 28
  homology 15, 16, 34
family cutoff value 94
family homology 122, 123
family table 115
family-clustering threshold 6
FAMILY-IDENTITY field 105
FAMILY-SCORE field 104
FASTA 4, 5, 9, 12, 41, 94, 110, 119
  homology 173
  NR homology EERIE 173
  SwissProt homology EERIE 173
FEATURE field 132
feature table 4, 144
Filter field 84, 87
Filter types
  internal, external 84
Filters
  multiple 84
Format for path 80
Full workup 23
functional motifs 11, 25, 44, 138

G
Gap spanning 110
GAP-CHARS field 105
gaps 30
GenBank 4, 7, 136
genetic map/linkage 14
GenPept NR BLAST homology 163
GenQuest
  BLAST (PDB) 167
  BLAST (Swiss) 167
  BLAST homology 163
  search (Netscape) 158
GET query 90
GLOBAL-SECURITY setting 65
GORF server 16, 141
GQ FASTA homology 174

H
Hide All Annotations 26, 27, 40
Highlight... command 49
history
  stamp 69
History field 79
history stamp 79
HOMOL field 109
Homolog 112
Homolog alignment 100
Homolog cluster 101
Homolog hits 94
Homolog mapping 112
Homolog scoring 102
  zero score truncation 103
homology
  annotation 5, 43
  annotation, in ORF 17
### Index

<table>
<thead>
<tr>
<th>Core Homology</th>
<th>Kringle</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>from BLAST/FASTA 48</td>
</tr>
<tr>
<td>hits 36</td>
<td>from Prosite 47</td>
</tr>
<tr>
<td>homology annotation, in ORF 141</td>
<td>from SCOP 48</td>
</tr>
<tr>
<td>Homology Fingerprint 93, 122</td>
<td>Kringle Domain motif 47</td>
</tr>
<tr>
<td>Homology fingerprint 15</td>
<td></td>
</tr>
<tr>
<td>homology fingerprint annotations 46</td>
<td></td>
</tr>
<tr>
<td>Homology information 161</td>
<td></td>
</tr>
<tr>
<td>homology leader sequence 101</td>
<td></td>
</tr>
<tr>
<td>Homology modeling 80</td>
<td></td>
</tr>
<tr>
<td>Homology Search 81</td>
<td></td>
</tr>
<tr>
<td>hotlist annotations 18</td>
<td></td>
</tr>
<tr>
<td>hotlist.hot 18</td>
<td></td>
</tr>
<tr>
<td>hyperlink 4, 30, 33</td>
<td></td>
</tr>
<tr>
<td>hypernote 30</td>
<td></td>
</tr>
</tbody>
</table>

**I**

- IDENTITY field 108
- identity scores in homolog alignment 100
- Inbox 35, 52, 66
- individual annotation 41
- INFOLINK field 107, 133
- information origins 1
- Information Window 51
- init1, 2 scores 103
- Input field 81
- internet 2
- intranet 2

**J**

- J. Biol. Chem Full Text 189
- JBC Online 189
- job groups 66
- job status report 67

**K**

- keyword 35
- keyword query 2
- Keyword Search 80
- Lod score 3, 5, 94
- Log file 67
- look 67
- LOOK_INSTALL directory 76
- look_output.log file 67

**L**

- LABEL field 108, 132
- Label field 87
- licensing 2
- Light workup 22, 24
- Link to annotation... 31
- links 31, 34, 49
- Literature icon 144
- Load Sequence... 38, 53
- lod score 3, 5, 94
- Log file 67
- look 67
- MEDLINE References 189
- MEMORY requirements 72
- Metabolic pathway/enzyme activity 138
- metabolic pathway/enzyme activity 14, 138
- MHC peptides database 177
- MINIMUM-SCORE field 104
- MISMATCH-CHARS field 105
- modeling 11, 116
- Molecule 81
- motif 11, 133
- Mouse modes 2, 59
- Multi-gene cluster 122
- multiple queries 68
- multiple services 79
Multiple sources 69
MultipleSeq Input variable 81
multi-sequence clusters 117
Mutant modeling 80
Mutation Annotations 114
Mutation annotations 116

N
Name field 79
NCBI 16, 141
ORF finder 179
network
  cache 66, 70
  query process 66
  sockets 66
new sequence 24
NIH grants 191
No workup 22
no-cache option 83
non-obvious search string 82
non-standard operator 82
NR BLAST homology 166
NR BLAST2 Homology 96
null hypothesis 5

O
OMIM 18, 144
open reading frame 16, 140
Open reading frames detection 177
ORF 57
ORFs (Any, >25aa) 178
ORFs (Met, >100aa) 178
Output field 81, 87
Output processing 96
overall pair identity 112
OWL 4
OWL database 158

P
pair-homology 112
papers 49
papers about annotations 49
parsing language 71
PATH defininitio 68
Path field 80, 87
Pattern field 86
PCR primers 17, 179
PCR scoring 143
PDB
  BLAST homology annotations 165
  database 147, 148
  hit and structural homology 11
  service 64
PDB records
  and SSP info 134
percent identity
  average of cluster 117
  low score against query 114
Perl script language 71
PFAM 133
PFAM pattern searching 11
Pick This 30
PID (UNIX process ID) 66
PIR
  annotations 152
  database 147, 151
  homology alignment 151
POST query 91
post-translational modification 11, 45
Prerelease PDB 150
Primary databases 148
primertx program 142
printf call 84
privacy level 25, 38, 39, 54
  changing in mid-session 39
PRODOM 4, 7, 136
  annotations 176
  database 176
ProDom database 136
Index

proprietary sequences
  release to public servers 25
PROSITE 4, 11, 44, 133
  auto annotations at EBI 181
  auto annotations at ExPASy 182
  protein motifs 180
Prosite text retrieval at ExPASy 181
PSSP
  prediction server 184
  report server 184
PSSP accuracy 135
Pubmed 18, 144
putative loop 9, 10, 43, 111
p-value 5, 94, 104
Python script language 71

Q
Queries
  maximum length 72
  simultaneous 66
query automation 68
QUERY field 109
query result 4
query types 64
QUERY-POS field 108
Queuing 66

R
RANGE field 133
Raw homology alignment 96
Raw homology score 94
re 39
READING-FRAME field 108
re-annotate sequence 39
recognition filter 71
Regular workup 22, 24
restriction site 16

S
SBJCT field 109
scanf call 84
SCOP 2, 4, 11, 44, 116
  BLAST 163
    BLAST homology annotations 166
    derivative of PDB 150
    superfamily 11
SCOP BLAST 94, 119
SCOP superfamily 101, 116, 118, 119
SCOP-ID field 108
SCORE field 108, 133
script-based filters 68
Search by service 91
Search My Documents 68
search result 35
secondary structure
  annotations 42
  prediction 8
Security
  assessment of 65
security levels 25, 39, 65
security preferences 1
Security values 65
SECURITY= expression 65
Seq input variable 81
SeqDBAnnotation 132
SeqDBAnnotation field 103
SeqDBAnnotation fields 132
SEQ-ID field 108
SEQREF-FORMAT 106
sequence
  alignments 12, 119
  databases 35
  homolog 9, 11, 43, 116
  homology 6
  identity 6
  numbering 45
  polymorphism 9, 111
  query 2
Sequence Analysis Refs 191
Index

Sequence Annotation 81
Sequence Databases 35, 159
Sequence homolog 110
sequence icon cluster 144
Sequence length constraint 72
Sequence polymorphic regions 114
Sequence Polymorphism 93
Sequence Queries 4, 13, 21, 39, 40, 80, 121
Sequence record 37
Servers
  upper limits on lengths 72
service groups 77
service type 78
SERVICE variable 77
Set Annotation Styles 29
Show Sequences 4, 67
significance cut-off 94
significance of family 100
SIMILAR-CHARS field 105
simultaneous queries 66
six-frame translation search 17, 141
Slow queries 70
Smith-Waterman 5
Snapshot 30
Snapshot option list 31
SOOP fields 90
SOURCE definition 68, 89
SOURCE entry 78
SOURCE field 78
SOURCE= variable 76
Sources 64
Specificity Pocket 137
specificity pocket 10, 137
sscanf match 86
state tracking in homolog mapping 112
STRAND-DIR field 108
Strip field 83
structural homolog 11
  and PDB hits 11
Structural homolog annotations 113
Structural Homology 93
structural homology 47
structural homology annotations 44
SUBJECT-POS field 109
Swiss FASTA homologies 119
SwissProt 7, 44, 45, 137
  Annotations 152
database 147, 152
SwissProt FASTA 94
System Command 91

T
Threshold cutoff 121
Title field 79
transcription factors 185
Translate in 6 frames 186
translated ORF 57
turquoise line
  in sequence homology 9, 114

U
U.S. Patents 192
UI field 80
UI=interface-type 80
uncertain annotation
  italicized 5
uncertainty 8, 134
UNIX 13, 15, 19, 120, 139
UNIX process identifier 66
URL checker 69
URL format 68
user-created annotation 20, 30, 33

V
Validation 124
very close sequence identity 114
View Highlight sub-menu 50
Viewer field 83
Virtual document 68
Index

W
WebCutter 4
WebCutter unique restriction sites 16, 24, 140
Webcutter unique restriction sites 182
white line
   in sequence homology 9, 114
Whitespace field 82
Wildcard field 82
workup level 4, 20, 22, 27, 38, 39, 54, 56, 58
   close homologs vs. new sequences 24
   control of 24
   full 23
   global control 24
   light 22
   none 22
   regular 22
Worldbase 1, 13, 18, 25, 75, 121
   customization 147
   in /home directory 76
   in /local directory 76
   organization 147
   source definitions 78
   system wide customization 76
   user customization 76
Worldbase application definitions 85
Worldbase customization 75
Worldbase event definitions 88
Worldbase format 77
Worldbase services 64, 77
worldbase.src 3, 22, 25, 69, 71
   multiple copies of 76
WWW buffer 70

X
xprimer Algorithm 143
xprimer program 142