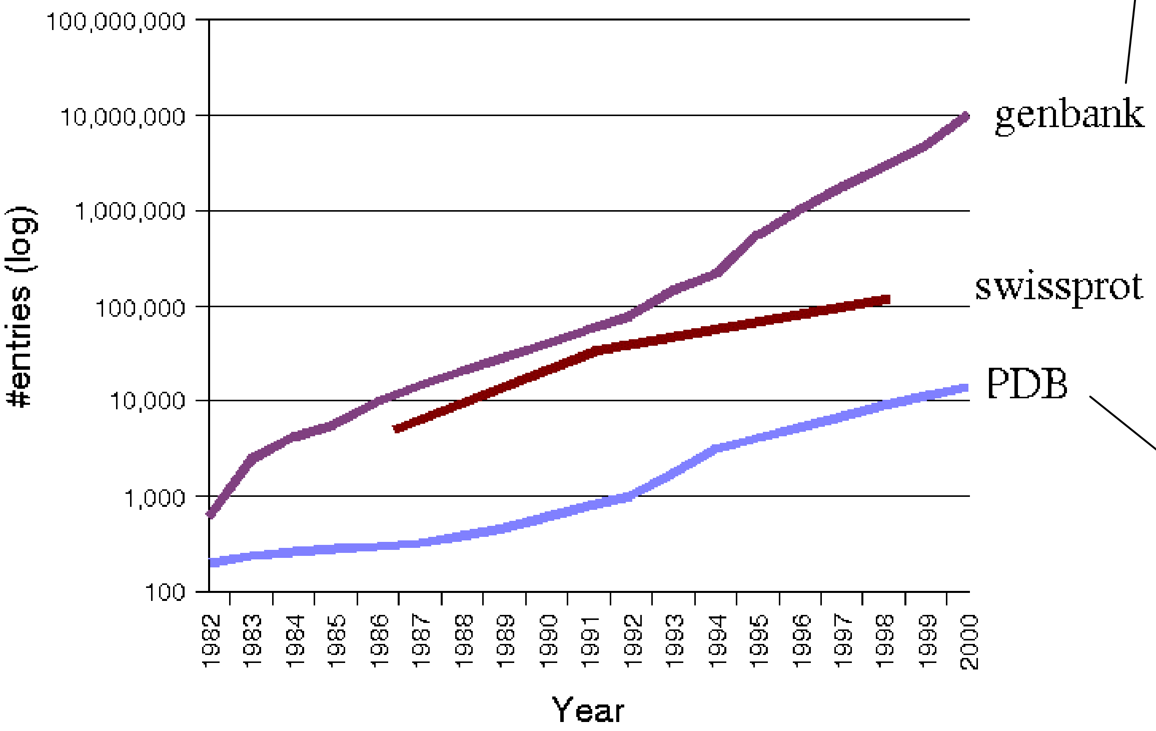
Protein Sequence Analysis

Exploiting Sequence Relationships for Predicting Protein Function

Florencio Pazos (CNB-CSIC)

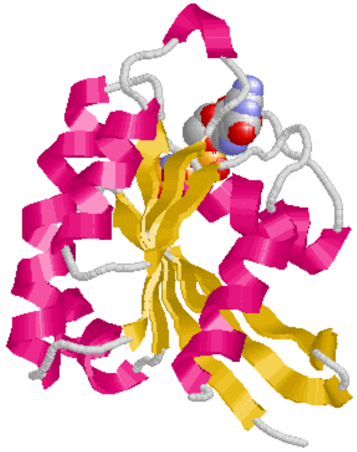
Florencio Pazos Cabaleiro
Protein Design Group (CNB-CSIC)
pazos@cnb.uam.es

Protein Sequences, Structures and Functions



MREYKLVVLGSGGVGKSALTVQFVQGIFVDEYDPTIEDSY
RKQVEVDCQQCMLEILDTAGTEQFTAMRDLYMKNGQGFAL
VYSITAQSTFNDLQDLREQILRVKDTEDVPMILVGNKCDL
EDERVVGKEQGQNLARQWCNCAFLESSAKSKINVNEIFYD
LVRQINR

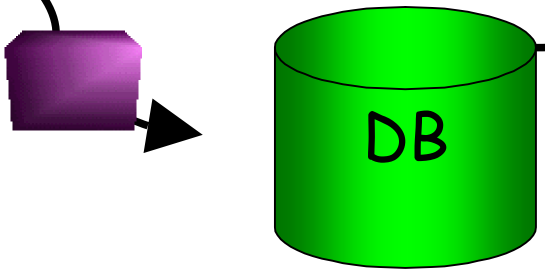
MLEILDTA**GTEQFT**AMRDLYMKNGQGFAL
VYSITAQSTFND**LQDLREQIL**RVKDTEDVPMIL
VGNKCDLEDERV

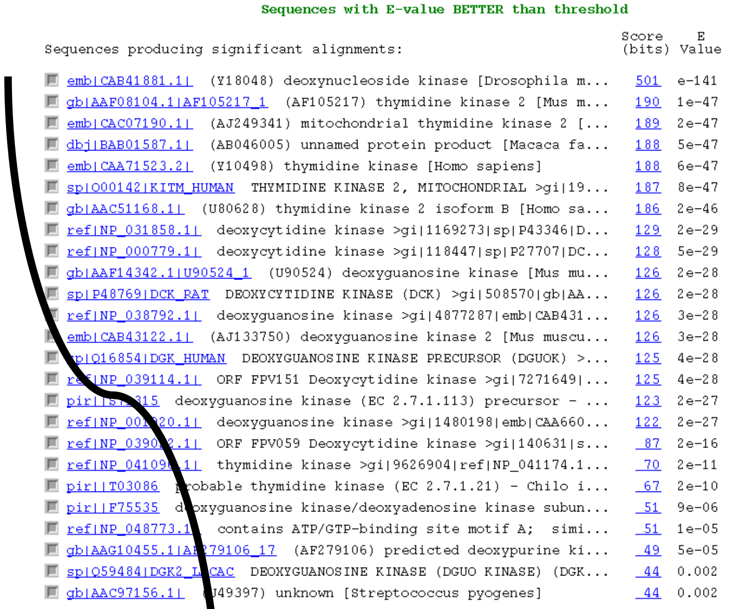


General strategy

NewSequence

Similarity search (BLAST, FASTA, ...)





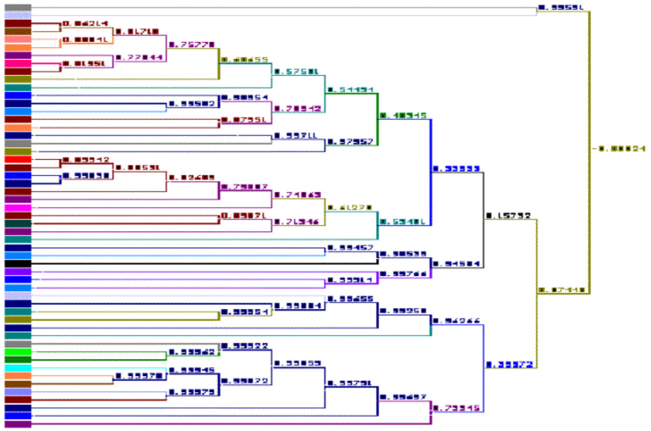
```
                          Sequences with E-value BETTER than threshold

                                                                            Score    E
Sequences producing significant alignments:                                (bits) Value

▣ emb|CAB41881.1|  (Y18048) deoxynucleoside kinase [Drosophila m...   501   e-141
▣ gb|AAF08104.1|AF105217_1  (AF105217) thymidine kinase 2 [Mus m...   190   1e-47
▣ emb|CAC07190.1|  (AJ249341) mitochondrial thymidine kinase 2 [...   189   2e-47
▣ dbj|BAB01587.1|  (AB046005) unnamed protein product [Macaca fa...   188   5e-47
▣ emb|CAA71523.2|  (Y10498) thymidine kinase [Homo sapiens]           188   6e-47
▣ sp|O00142|KITM_HUMAN  THYMIDINE KINASE 2, MITOCHONDRIAL >gi|19...   187   8e-47
▣ gb|AAC51168.1|  (U80628) thymidine kinase 2 isoform B [Homo sa...   186   2e-46
▣ ref|NP_031858.1|  deoxycytidine kinase >gi|1169273|sp|P43346|D...   129   2e-29
▣ ref|NP_000779.1|  deoxycytidine kinase >gi|118447|sp|P27707|DC...   128   5e-29
▣ gb|AAF14342.1|U90524_1  (U90524) deoxyguanosine kinase [Mus mu...   126   2e-28
▣ sp|P48769|DCK_RAT  DEOXYCYTIDINE KINASE (DCK) >gi|508570|gb|AA...   126   2e-28
▣ ref|NP_038792.1|  deoxyguanosine kinase >gi|4877287|emb|CAB431...   126   3e-28
▣ emb|CAB43122.1|  (AJ133750) deoxyguanosine kinase 2 [Mus muscu...   126   3e-28
▣ sp|Q16854|DGK_HUMAN  DEOXYGUANOSINE KINASE PRECURSOR (DGUOK) >...   125   4e-28
▣ ref|NP_039114.1|  ORF FPV151 Deoxycytidine kinase >gi|7271649|...   125   4e-28
▣ pir||S71315  deoxyguanosine kinase (EC 2.7.1.113) precursor - ...   123   2e-27
▣ ref|NP_001920.1|  deoxyguanosine kinase >gi|1480198|emb|CAA660...   122   2e-27
▣ ref|NP_039022.1|  ORF FPV059 Deoxycytidine kinase >gi|140631|s...    87   2e-16
▣ ref|NP_041096.1|  thymidine kinase >gi|9626904|ref|NP_041174.1...    70   2e-11
▣ pir||T03086  probable thymidine kinase (EC 2.7.1.21) - Chilo i...    67   2e-10
▣ pir||F75535  deoxyguanosine kinase/deoxyadenosine kinase subun...    51   9e-06
▣ ref|NP_048773.1|  contains ATP/GTP-binding site motif A;  simi...    51   1e-05
▣ gb|AAG10455.1|AF279106_17  (AF279106) predicted deoxypurine ki...    49   5e-05
▣ sp|Q59484|DGK2_LACAC  DEOXYGUANOSINE KINASE (DGUO KINASE) (DGK...    44   0.002
▣ gb|AAC97156.1|  (U49397) unknown [Streptococcus pyogenes]            44   0.002
```

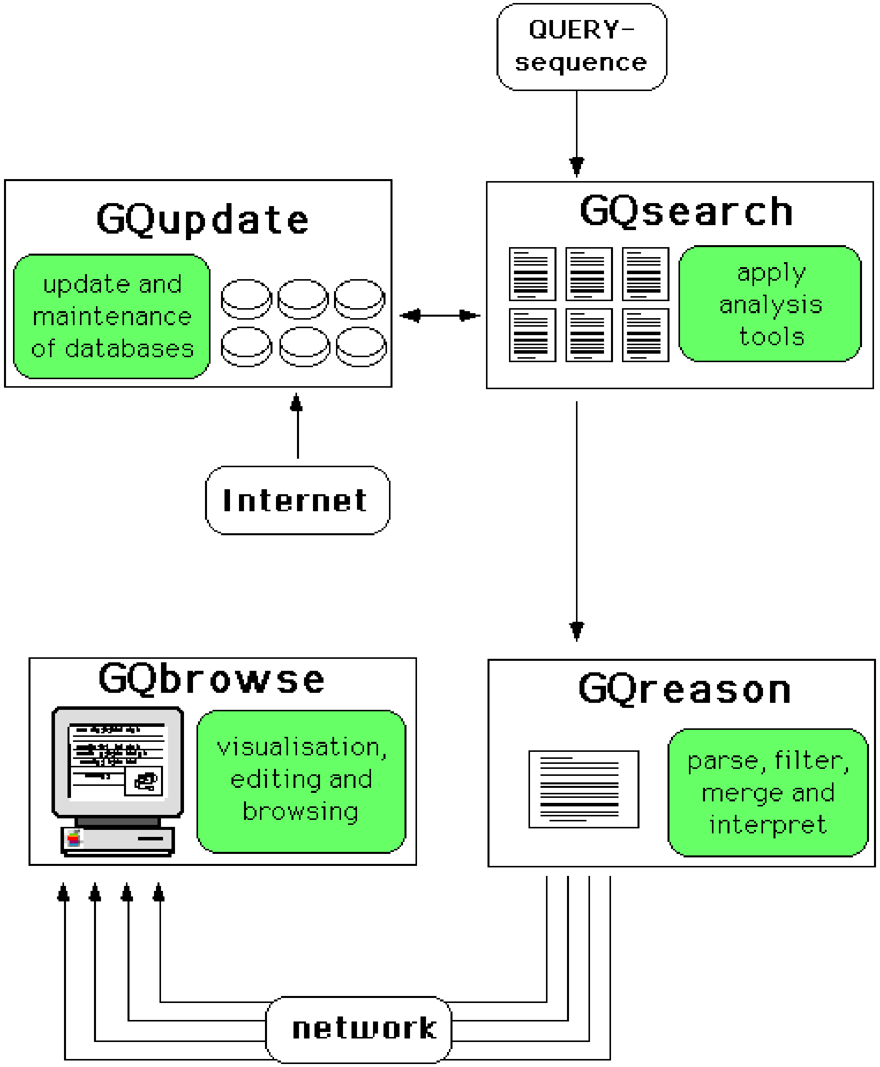
```
Score =  124 bits (309), Expect = 1e-27
 Identities = 108/340 (31%), Positives = 156/340 (45%), Gaps = 62/340 (18%)

Query: 1   ATTYNAVVSKSSSDGKTFKTIADAIASAPAGSTP-FVILIKNGVYNERLTITRN--NLHL 5
           + T NAVV+   S    FKT+A A+A+AP G T  ++I IK GVY E + +T++    N+
Sbjct: 269 SVTPNAVVAADGSGN--FKTVAAAVAAAPQGGTKRYIIRIKAGVYRENVEVTKKHKNI
```



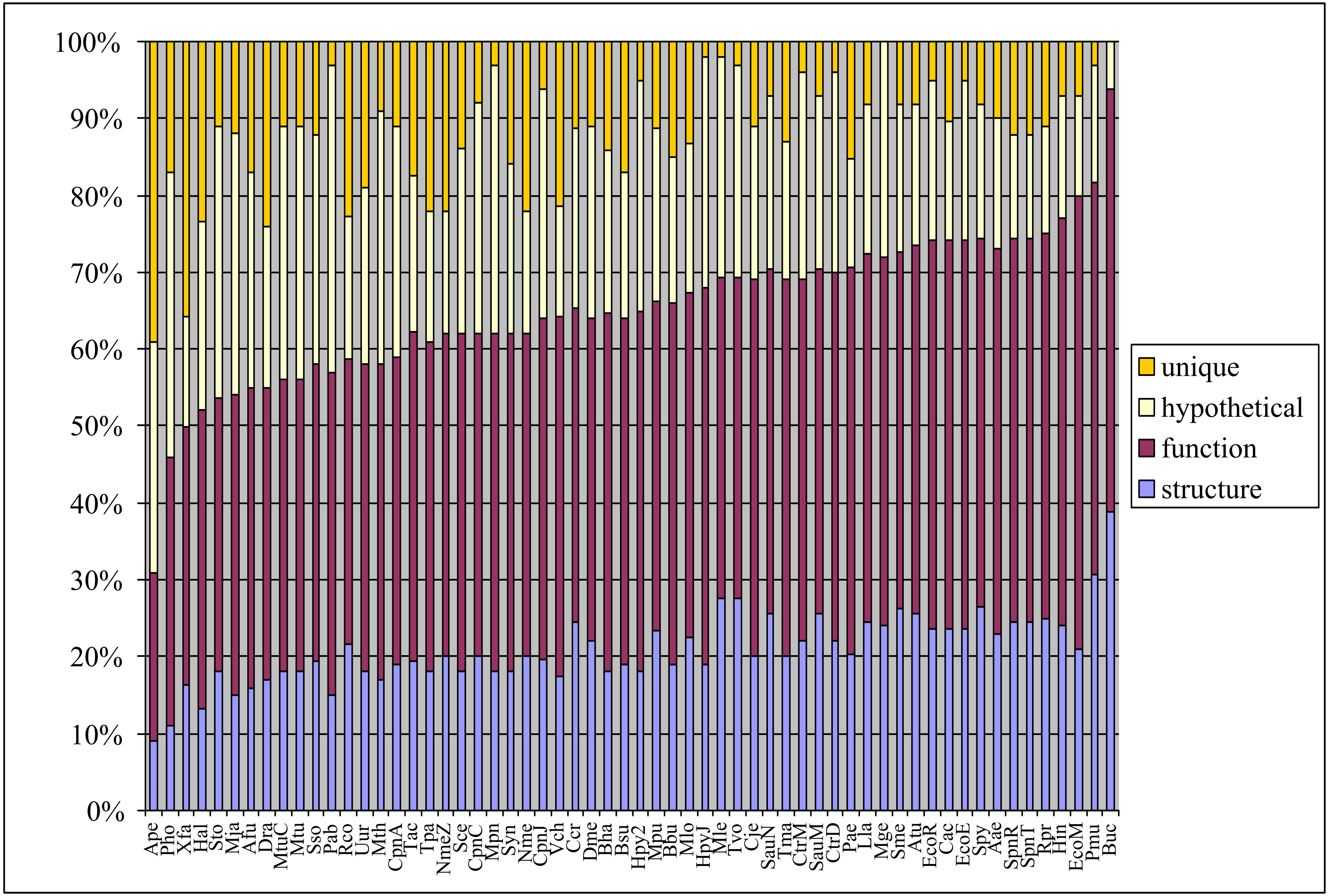
Deduced homology

High throughput application - *GeneQuiz*

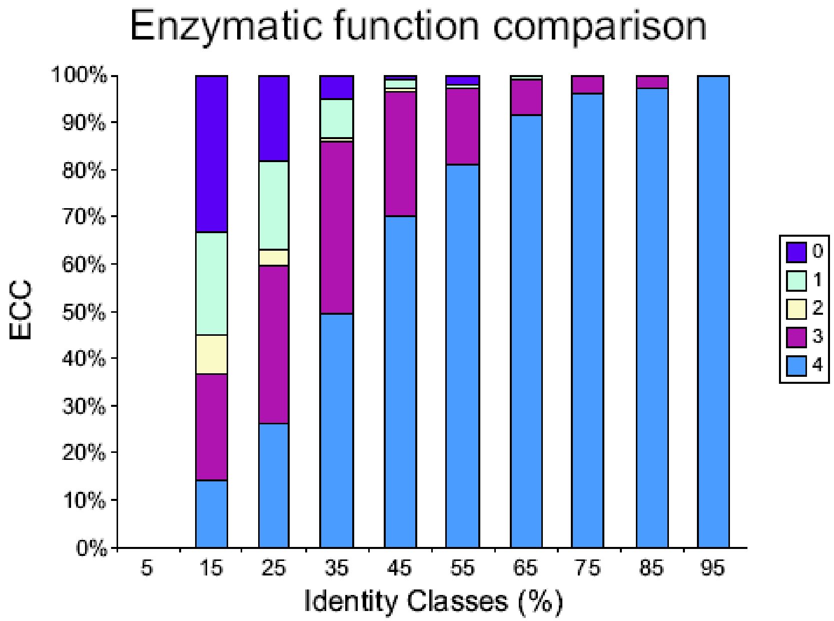

QUERY-sequence
GQupdate
update and maintenance of databases
GQsearch
apply analysis tools
Internet
GQbrowse
visualisation, editing and browsing
GQreason
parse, filter, merge and interpret
network

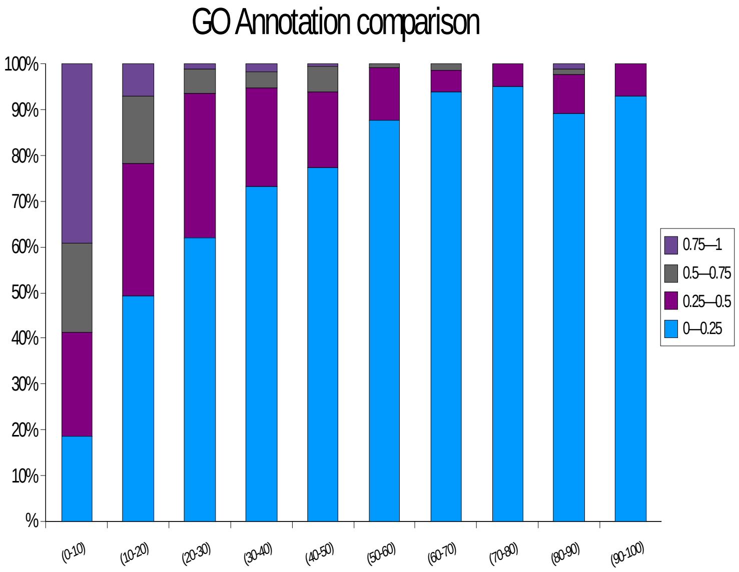
- Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C. and Sander, C. (1994) GeneQuiz: a workbench for sequence analysis. *Proc Int Conf Intell Syst Mol Biol.*, **2**, 348-353.
- Hoersch, S., Leroy, C., Brown, N.P., Andrade, M.A. and Sander, C. (2000) The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem Sci.*, **25**, 33-35.

High throughput application - *GeneQuiz*



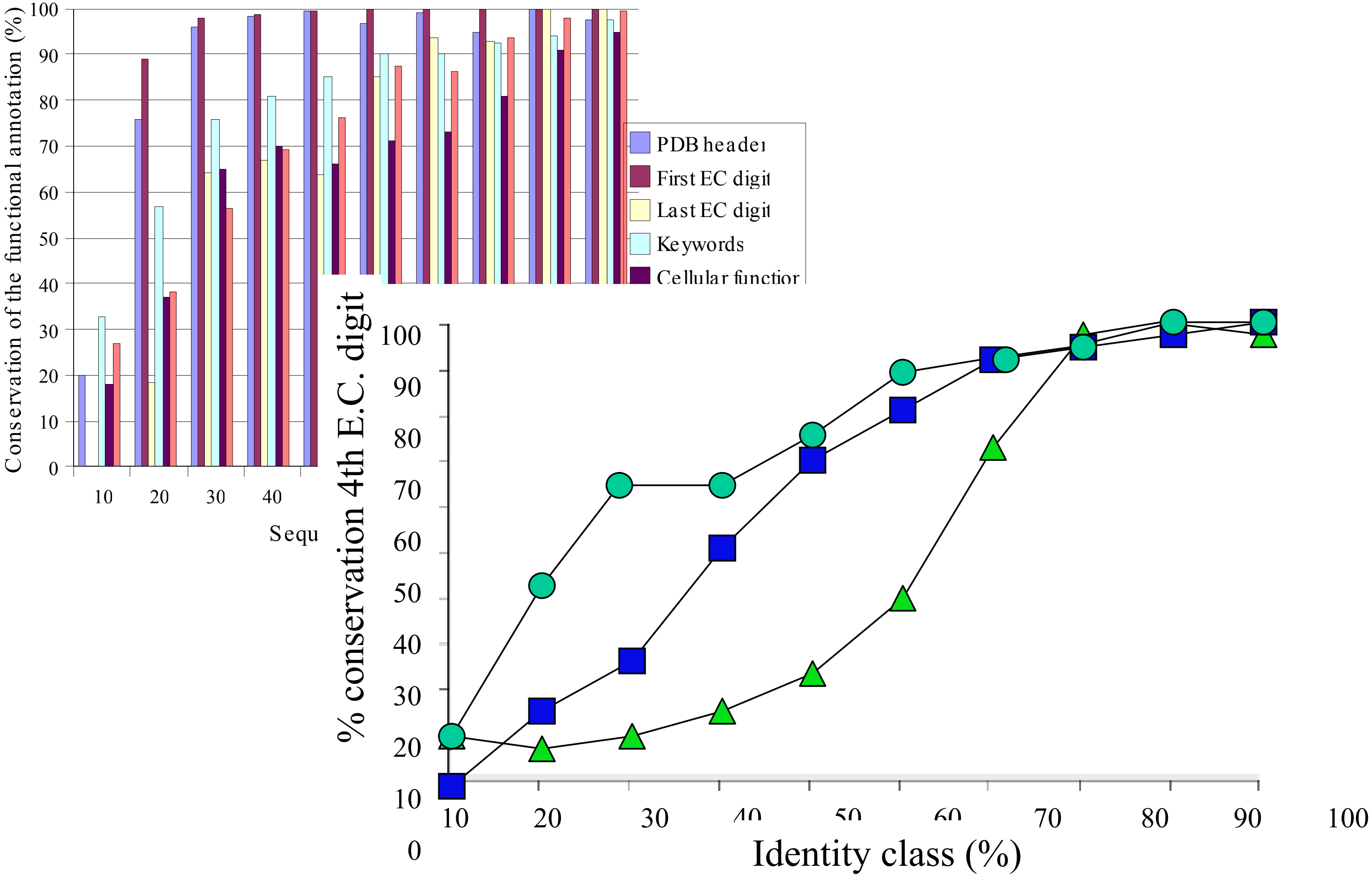
How reliable is the similarity-based Functional transfer?





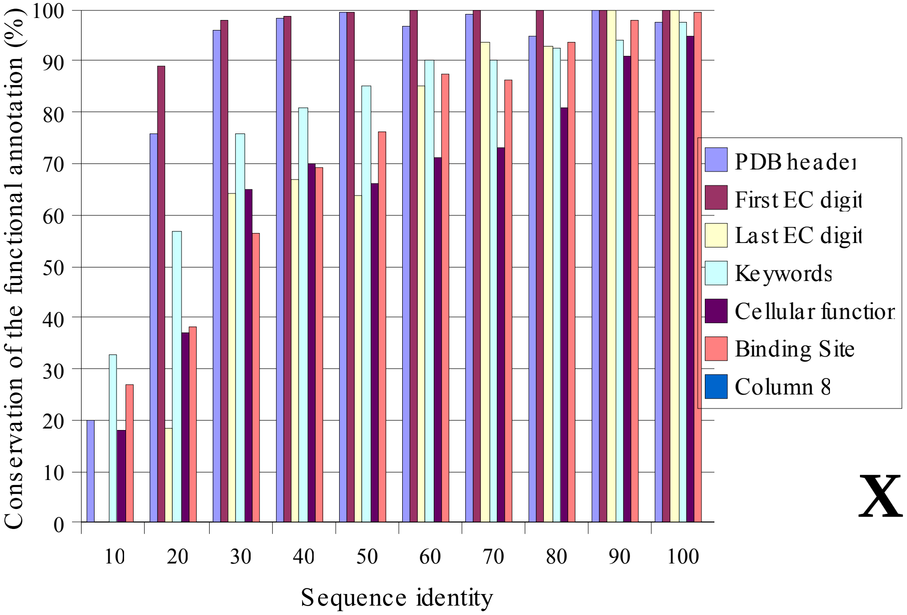
- Devos, D. and Valencia, A. (2000) Practical limits of function prediction. *Proteins*, **41**, 98-107.
- Valencia, A. (2005) Automatic annotation of protein function. *Curr Opin Struct Biol*, **15**, 267-274.

How reliable is the similarity-based Functional transfer?

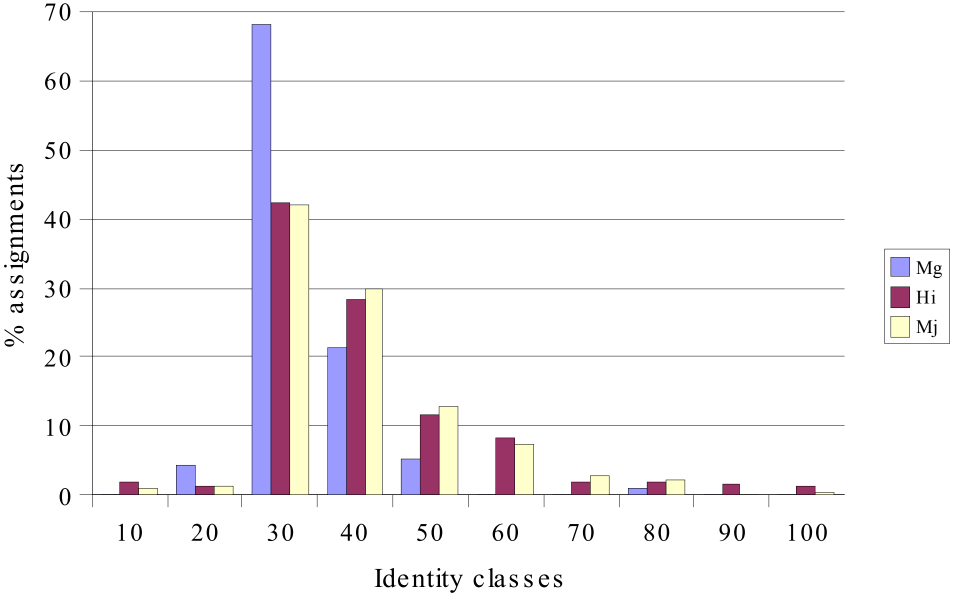


Valencia, A. (2005) Automatic annotation of protein function. *Curr Opin Struct Biol*, **15**, 267-274.

So… Which degree of error can we expect?



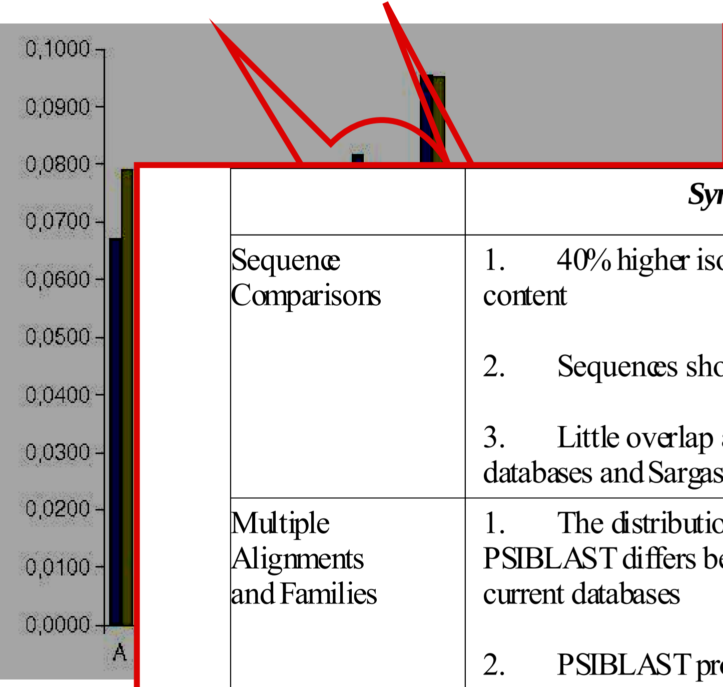
X



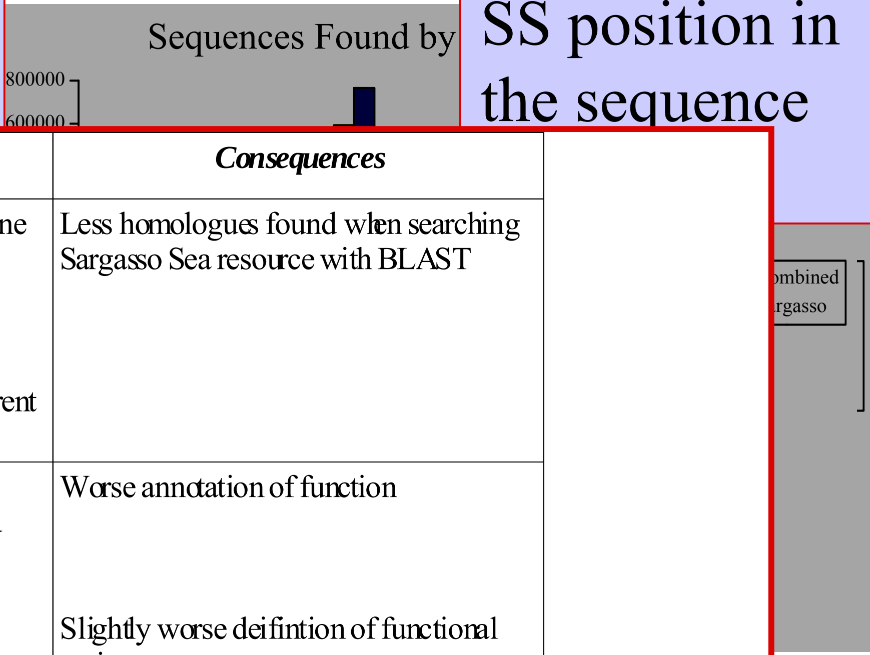
	PDB header	First EC Digit	Last EC Digit	Key-words	Funct. class	Binding site
Mg	4	2	37	23	35	40
Hi	4	2	31	20	33	34
Mj	4	2	32	20	33	34

Devos, D. and Valencia, A. (2001) Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429-431.

Environmental Genomics (Sargasso Sea)

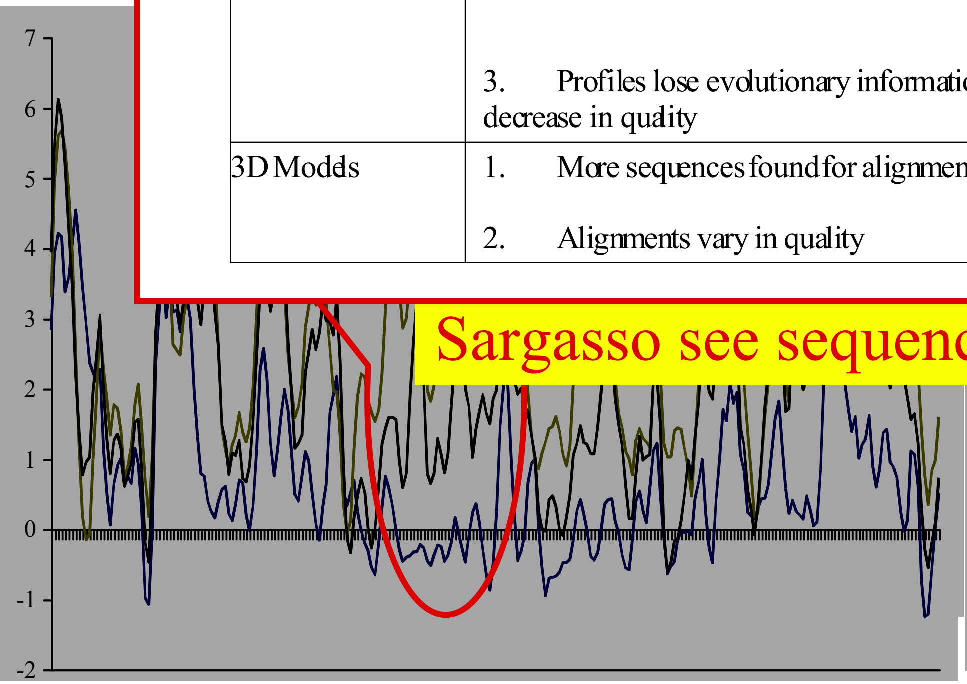


Differentia
1 sequence

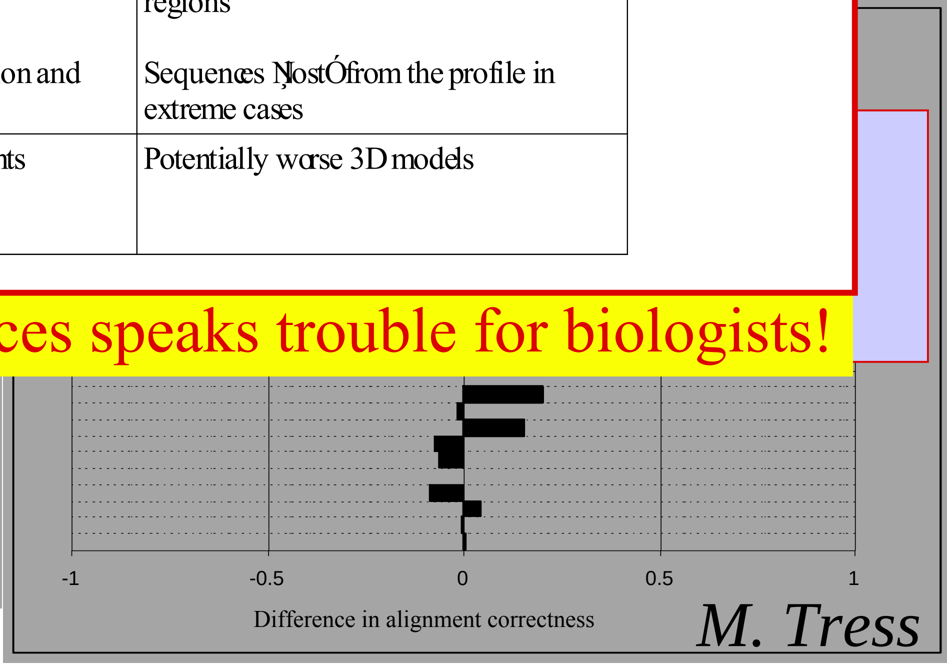


SS position in the sequence

	Symptoms	*Consequences*
Sequence Comparisons	1. 40% higher isoleucine, asparagine and lysine content 2. Sequences shorter and more fragments 3. Little overlap at 90% identity between current databases and Sargasso Sea	Less homologues found when searching Sargasso Sea resource with BLAST
Multiple Alignments and Families	1. The distribution of sequences found by PSIBLAST differs between the Sargasso Sea and current databases 2. PSIBLAST profiles drift more 3. Profiles lose evolutionary information and decrease in quality	Worse annotation of function Slightly worse deifintion of functional regions Sequences ÑlostÓ from the profile in extreme cases
3D Models	1. More sequences found for alignments 2. Alignments vary in quality	Potentially worse 3D models

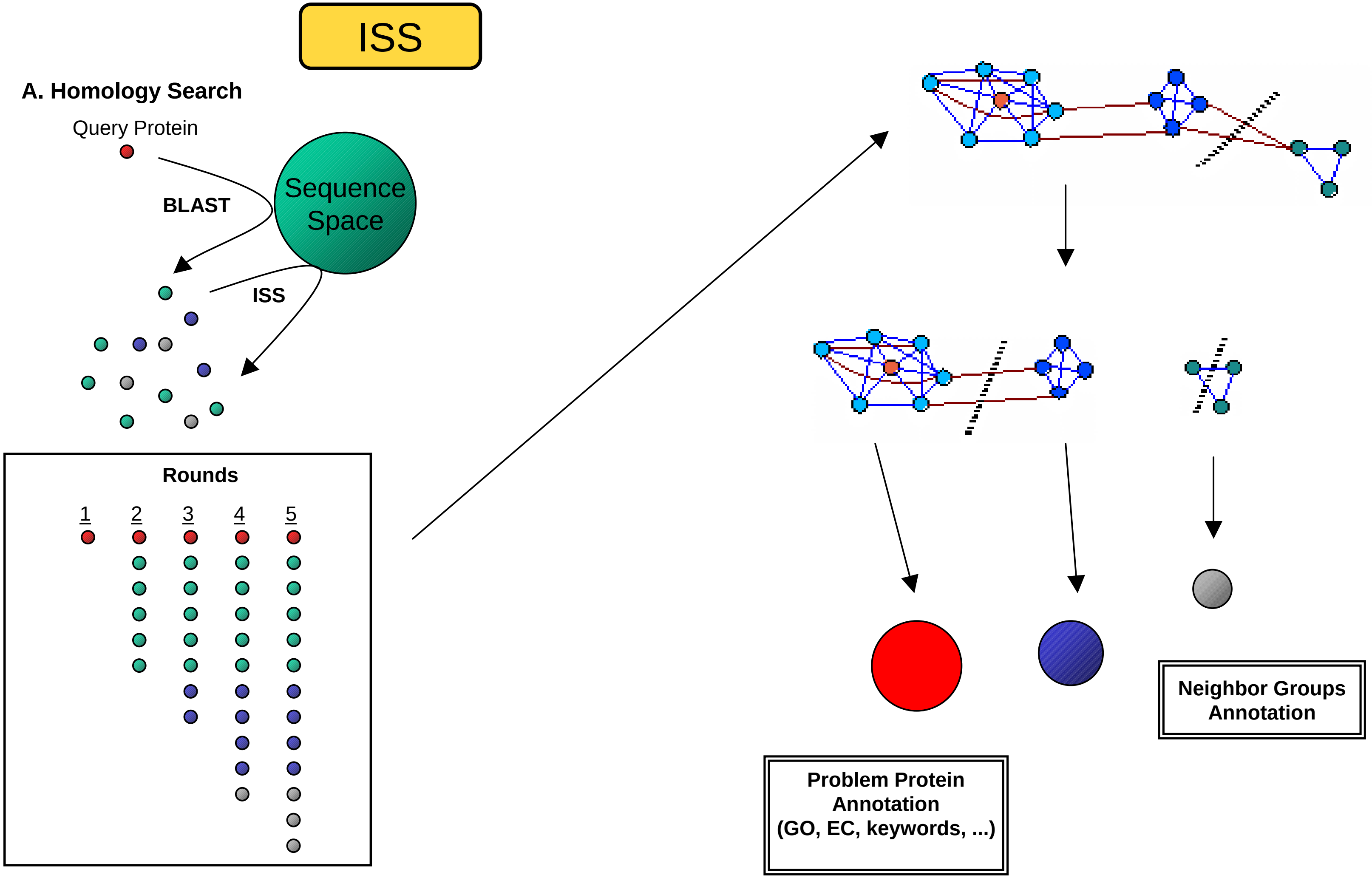


Sargasso see sequences speaks trouble for biologists!



M. Tress

More complex strategies
FunCut



ISS

A. Homology Search

Query Protein

BLAST

Sequence Space

ISS

Rounds

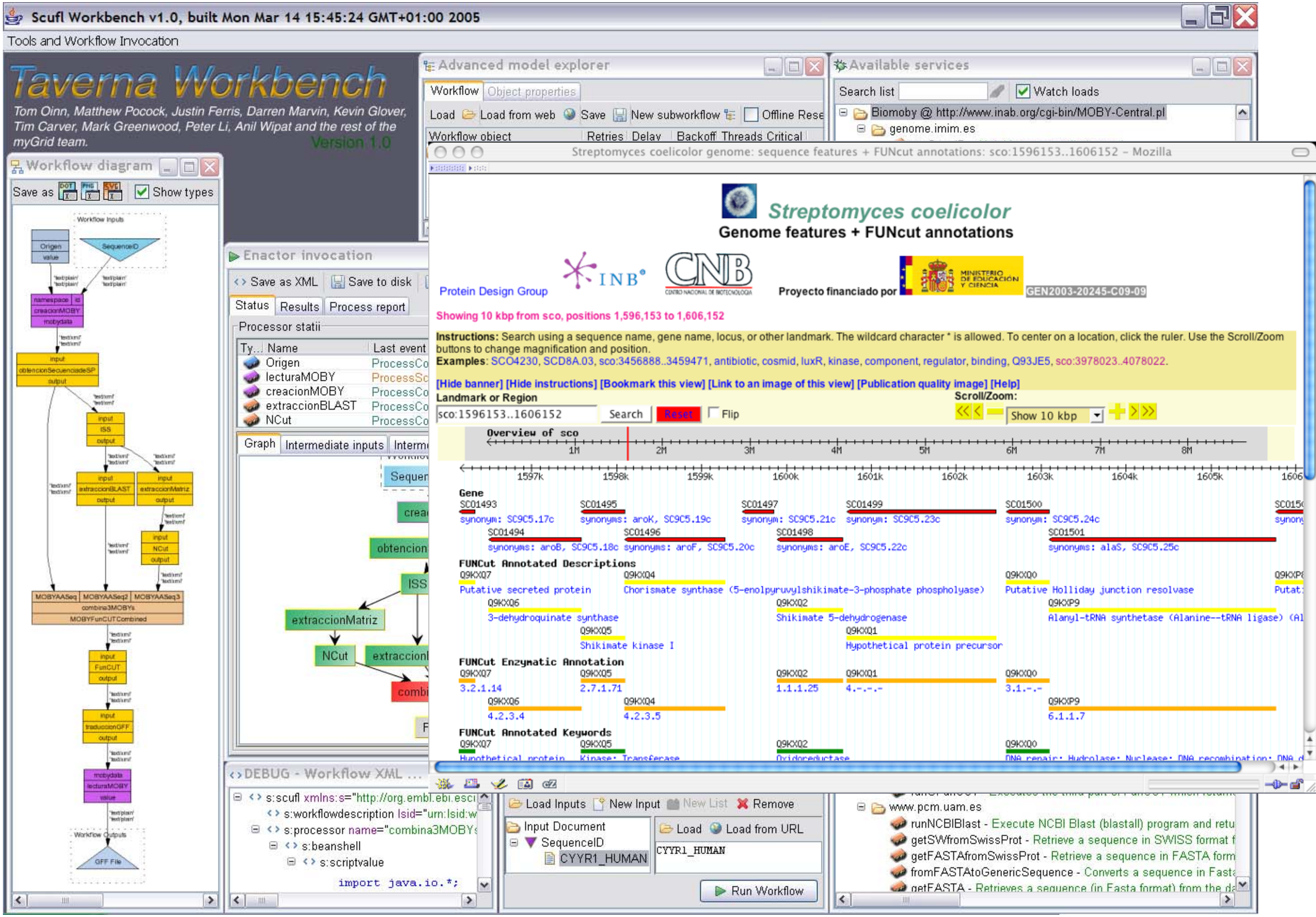
1 2 3 4 5

Neighbor Groups Annotation

Problem Protein Annotation (GO, EC, keywords, ...)

- Abascal, F. and Valencia, A. (2002) Clustering of proximal sequence space for the identification of protein families. *Bioinformatics*., **18**, 908-921.
- Abascal, F. and Valencia, A. (2003) Automatic annotation of protein function based on family identification. *Proteins*, **53**, 683-692.

Integrating annotation services



Tom Oinn, Matthew Pocock, Justin Ferris, Darren Marvin, Kevin Glover, Tim Carver, Mark Greenwood, Peter Li, Anil Wipat and the rest of the myGrid team.

Processor statii

Ty...	Name	Last event
	Origen	ProcessCo
	lecturaMOBY	ProcessSc
	creacionMOBY	ProcessCo
	extraccionBLAST	ProcessCo
	NCut	ProcessCo

```
<> s:scufl xmlns:s="http://org.embl.ebi.escie
  <> s:workflowdescription lsid="urn:lsid:w
  <> s:processor name="combina3MOBYs
    <> s:beanshell
      <> s:scriptvalue
            import java.io.*;
```

Workflow object	Retries	Delay	Backoff	Threads	Critical

- Biomoby @ http://www.inab.org/cgi-bin/MOBY-Central.pl
  - genome.imim.es

- SequenceID
  - CYYR1_HUMAN

```
CYYR1_HUMAN
```

- www.pcm.uam.es
  - runNCBIBlast - Execute NCBI Blast (blastall) program and retu
  - getSWfromSwissProt - Retrieve a sequence in SWISS format f
  - getFASTAfromSwissProt - Retrieve a sequence in FASTA form
  - fromFASTAtoGenericSequence - Converts a sequence in Fasta
  - getFASTA - Retrieves a sequence (in Fasta format) from the da

Streptomyces coelicolor genome: sequence features + FUNcut annotations: sco:1596153..1606152 - Mozilla

Streptomyces coelicolor

Genome features + FUNcut annotations

Protein Design Group INB CNB CENTRO NACIONAL DE BIOTECNOLOGIA **Proyecto financiado por** MINISTERIO DE EDUCACIÓN Y CIENCIA GEN2003-20245-C09-09

Showing 10 kbp from sco, positions 1,596,153 to 1,606,152

Instructions: Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: SCO4230, SCD8A.03, sco:3456888..3459471, antibiotic, cosmid, luxR, kinase, component, regulator, binding, Q93JE5, sco:3978023..4078022.

[Hide banner] [Hide instructions] [Bookmark this view] [Link to an image of this view] [Publication quality image] [Help]

Landmark or Region sco:1596153..1606152 Search Reset □ Flip

Scroll/Zoom: Show 10 kbp

Gene

SC01493 synonym: SC9C5.17c
SC01494 synonyms: aroB, SC9C5.18c
SC01495 synonyms: aroK, SC9C5.19c
SC01496 synonyms: aroF, SC9C5.20c
SC01497 synonym: SC9C5.21c
SC01498 synonyms: aroE, SC9C5.22c
SC01499 synonym: SC9C5.23c
SC01500 synonym: SC9C5.24c
SC01501 synonyms: alaS, SC9C5.25c
SC015 synony

FUNCut Annotated Descriptions

Q9KXQ7 Putative secreted protein
Q9KXQ6 3-dehydroquinate synthase
Q9KXQ5 Shikimate kinase I
Q9KXQ4 Chorismate synthase (5-enolpyruvylshikimate-3-phosphate phospholyase)
Q9KXQ2 Shikimate 5-dehydrogenase
Q9KXQ1 Hypothetical protein precursor
Q9KXQ0 Putative Holliday junction resolvase
Q9KXP9 Alanyl-tRNA synthetase (Alanine--tRNA ligase) (Al
Q9KXP Putat

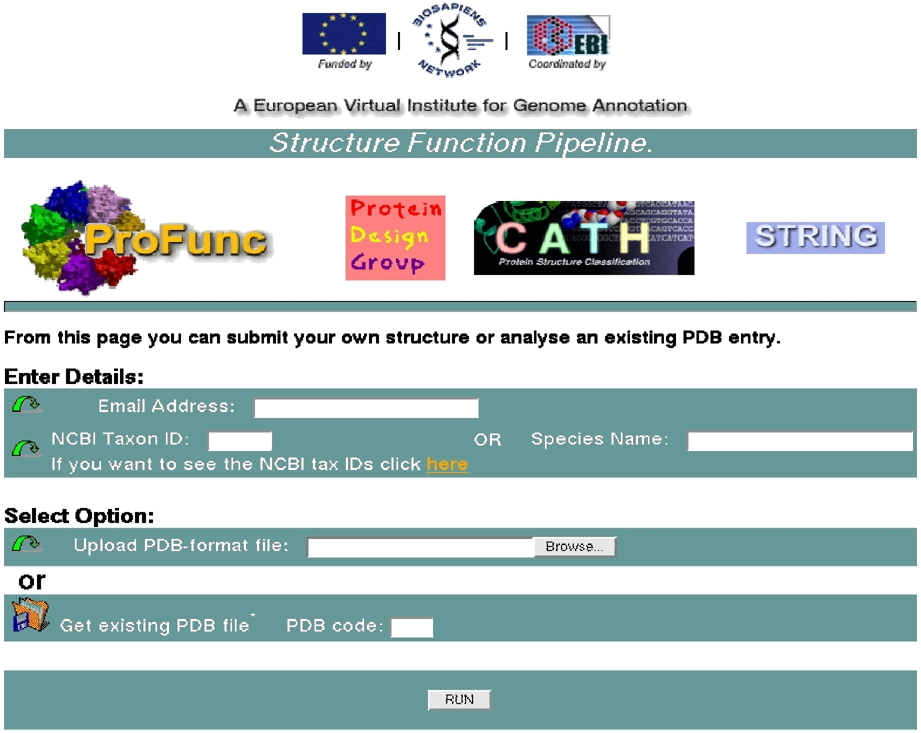
FUNCut Enzymatic Annotation

Q9KXQ7 3.2.1.14
Q9KXQ6 4.2.3.4
Q9KXQ5 2.7.1.71
Q9KXQ4 4.2.3.5
Q9KXQ2 1.1.1.25
Q9KXQ1 4.-.-.-
Q9KXQ0 3.1.-.-
Q9KXP9 6.1.1.7

FUNCut Annotated Keywords

Q9KXQ7 Hypothetical protein
Q9KXQ5 Kinase; Transferase
Q9KXQ2 Oxidoreductase
Q9KXQ0 DNA repair; Hydrolase; Nuclease; DNA recombination; DNA d

Integrating annotation services



Funded by | Coordinated by

A European Virtual Institute for Genome Annotation

Structure Function Pipeline.

ProFunc | Protein Design Group | CATH Protein Structure Classification | STRING

From this page you can submit your own structure or analyse an existing PDB entry.

Enter Details:

Email Address:

NCBI Taxon ID: OR Species Name:

If you want to see the NCBI tax IDs click here

Select Option:

Upload PDB-format file: Browse...

or

Get existing PDB file PDB code:

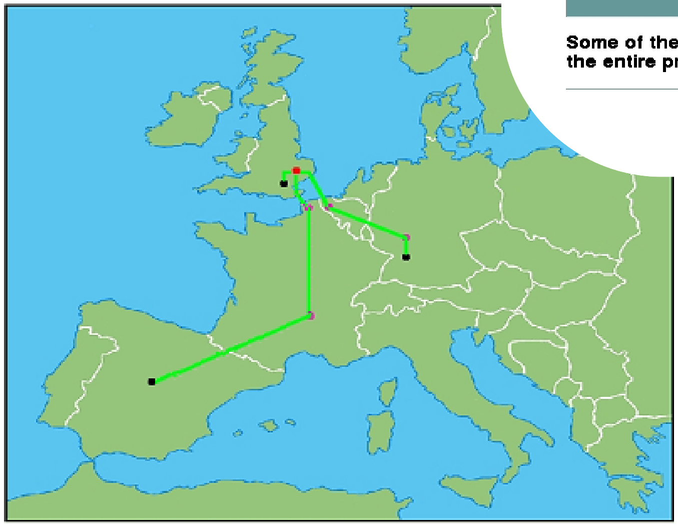
RUN

Some of the methods take minutes to run; others take hours. You will be notified by e-mail when the entire process is complete, but can check on preliminary results as they become available.

E-mail to:- *James Watson* *watson@ebi.ac.uk*

Key to map:

QUEUED	RUNNING	COMPLETED	FAIL
■	🟥	🟩	🟪



•Utilises SOAP interfaces to simultaneously access:

- **ProFunc (EBI, Hinxton, UK)**
- **CATH (UCL, London, UK)**
- **FUNcut (CNB, Madrid, Spain)**
- **STRING (EMBL, Heidelberg, Germany)**