Software

A comparative proteomics resource: proteins of *Arabidopsis thaliana*

Wilfred W Li*, Greg B Quinn*, Nickolai N Alexandrov†, Philip E Bourne*‡ and Ilya N Shindyalov*

Addresses: *San Diego Supercomputer Center, 9500 Gilman Drive, University of California San Diego, La Jolla, CA 92093-0505, USA. †Ceres Inc., 3007 Malibu Canyon Road, Malibu, CA 90265, USA. ‡Department of Pharmacology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA.

Correspondence: Philip E Bourne. E-mail: bourne@sdsc.edu

Published: 28 July 2003

Genome ***Biology*** 2003, **4:**R51

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/8/R51

Received: 3 February 2003
Revised: 6 May 2003
Accepted: 2 July 2003

Abstract

Using an integrative genome annotation pipeline (iGAP) for proteome-wide protein structure and functional domain assignment, we analyzed all the proteins of *Arabidopsis thaliana*. Three-dimensional structures at the level of the domain are assigned by fold recognition and threading based on a novel fold library that extends common domain classifications. iGAP is being applied to proteins from all available proteomes as part of a comparative proteomics resource. The database is accessible from the web.

Rationale

Protein-sequence-based comparative analysis to infer biological function is important and familiar to most biologists. Sequence-profile methods such as PSI-BLAST [1] or HMMER [2] are often used to detect distant homologs, and resources such as Prosite [3], BLOCKS [4] and PFAM [5] are representative resources resulting from protein classification based on sequence patterns. Protein structure also plays a crucial role in a full understanding of protein function as it is more conserved than sequence and hence exposes relationships not possible from sequence alone. Many protein domains have less than 10% sequence identity, and yet possess a similar fold and possibly related function.

One of the early insights gained from comparative genomics was domain accretion [6]. From prokaryotes to eukaryotes, the number of domains increases. But in higher eukaryotes, different combinations of domains are often observed in the same and different protein families. From a structural point of view domains are discreet compact folding units. PIR [7] classifies proteins into either a homeomorphic superfamily (proteins containing similar domains in the same order) or a homology domain superfamily (proteins from different homeomorphic superfamilies sharing a common ancestral domain). This modular nature of proteins necessitates a new approach to proteome annotation - a structural-domain-based approach.

There already exist a number of automated or semi-automated complete genome annotation systems. For example, GeneQuiz [8] and PEDANT [9] are two pipelines that are comprehensive and highly automated (Table 1). Similarly, there are several sites that provide protein structure annotations for various genomes. Superfamily [10] uses a set of hidden Markov model (HHM) profiles based on SCOP superfamily members. MatDB, based on PEDANT analysis of *Arabidopsis thaliana*, provides structural annotations using SCOP domain position specific scoring matrix (PSSM) profiles. The National Center for Biotechnology Information (NCBI) maintains a Conserved Domain Database (CDD) that

Table 1

Comparison of different annotation pipelines

Pipeline	Focus area	Applications	Coverage
GeneQuiz	Sequence homology Function assignment	BLAST, FASTA, COILS, MaxHom, Prosite, Blocks, Predict Protein, Coils, Transmembrane helix, CAST.	65 genomes
PEDANT	Gene prediction Sequence homology Function assignment Fold assignment	BLAST, PSI-BLAST, HMMER, PREDATOR, Orpheus, BLIMPS, STRIDE.	133 complete genomes, 91 partial genomes
PAT	Sequence homology Function assignment Fold recognition Structure prediction	WU-BLAST, PSI-BLAST, 123D, HMMER	103+ genomes, continuous expansion

uses PFAM and SMART [11] domain PSSMs to detect possible structural homologs. The 3D-Genomics database [12] uses SCOP domain PSSMs from 3D-PSSM [13]. Gene3D uses the CATH domain classification to annotate genes and genomes [14].

We have developed an automated integrative genome annotation pipeline (iGAP) initially to annotate the proteins of *A. thaliana* and later all proteomes based on a comprehensive fold library (Figure 1). In addition to the domains from SCOP, we have included domains parsed using the protein domain parser (PDP) [15], full-length Protein Data Bank (PDB) chains and chains not classified by SCOP, but associated with SCOP using combinatorial extension (CE), a structural-similarity search algorithm [16]. The result is a comprehensive fold library (FOLDLIB) from which comparative and fold recognition models of three-dimensional structure are derived. As a step beyond PSI-BLAST or PFAM profiles, we have used 123D+ [17,18], which not only performs target-template profile-profile alignment, but also uses secondary structure and contact capacity potential information for protein fold recognition. Further, the annotation pipeline provides a graded reliability index of functional prediction reliability ranging from A to E based on extensive benchmarking of selectivity versus sensitivity (N.N.A., I.N.S and P.E.B., unpublished work). Here we describe iGAP and the initial results on the analysis of *A. thaliana*, the first proteome processed, using a combination of web interface and SQL queries (Figure 2). Comparisons are made to other annotation schemes used to process *Arabidopsis* and to other proteomes processed with iGAP. The iGAP is systematically being applied to more than 1,000 proteomes, completely or partially sequenced and publicly available at NCBI [19], to develop a comparative proteomic resource.

Results and discussion

Automated annotation pipelines are crucial to organize the deluge of genomic information. Table 1 compares features of iGAP with those of GeneQuiz and PEDANT, two established genome annotation methodologies. GeneQuiz focuses on homolog and function assignment through sequence similarity search; PEDANT is a comprehensive analysis pipeline with emphasis on gene prediction, secondary and tertiary structure assignment; iGAP puts much more emphasis on fold recognition, threading and, to be released in the near future, homology modeling. Table 2 compares the proteins of *A. thaliana* (PAT) database to established databases of protein annotations. They differ in both coverage and focus. Again, each of the resources has clear strengths in a number of areas, but PAT stands out in terms of the amount of structural information it provides. Whereas other resources are limited to what is present in PDB or SCOP, PAT provides additional domains from PDP, and genetic domains from Astral. Moreover, an important feature of iGAP is the benchmarking used to establish the reliability measures. Such quality assurance is critical to the future development of these resources if they are to be used in a meaningful way by experimentalists.

Table 3a indicates the coverage of the *Arabidopsis* proteome provided by each methodology and associated resource. It is clear that InterPro and iGAP represent two approaches that provide very high coverage of the *Arabidopsis* proteome, based on sequence and structural information respectively. A combination of InterProScan and iGAP is under active development to integrate sequence- and structure-based annotation. Interestingly, only 14% of the *Arabidopsis* Information Resource (TAIR) GO annotation is based on nonelectronic annotation. This makes an even stronger argument for the integration of sequence- and structure-based annotation, to reduce the possibility of error propagation in electronic annotation. Table 3b highlights some specific examples of results

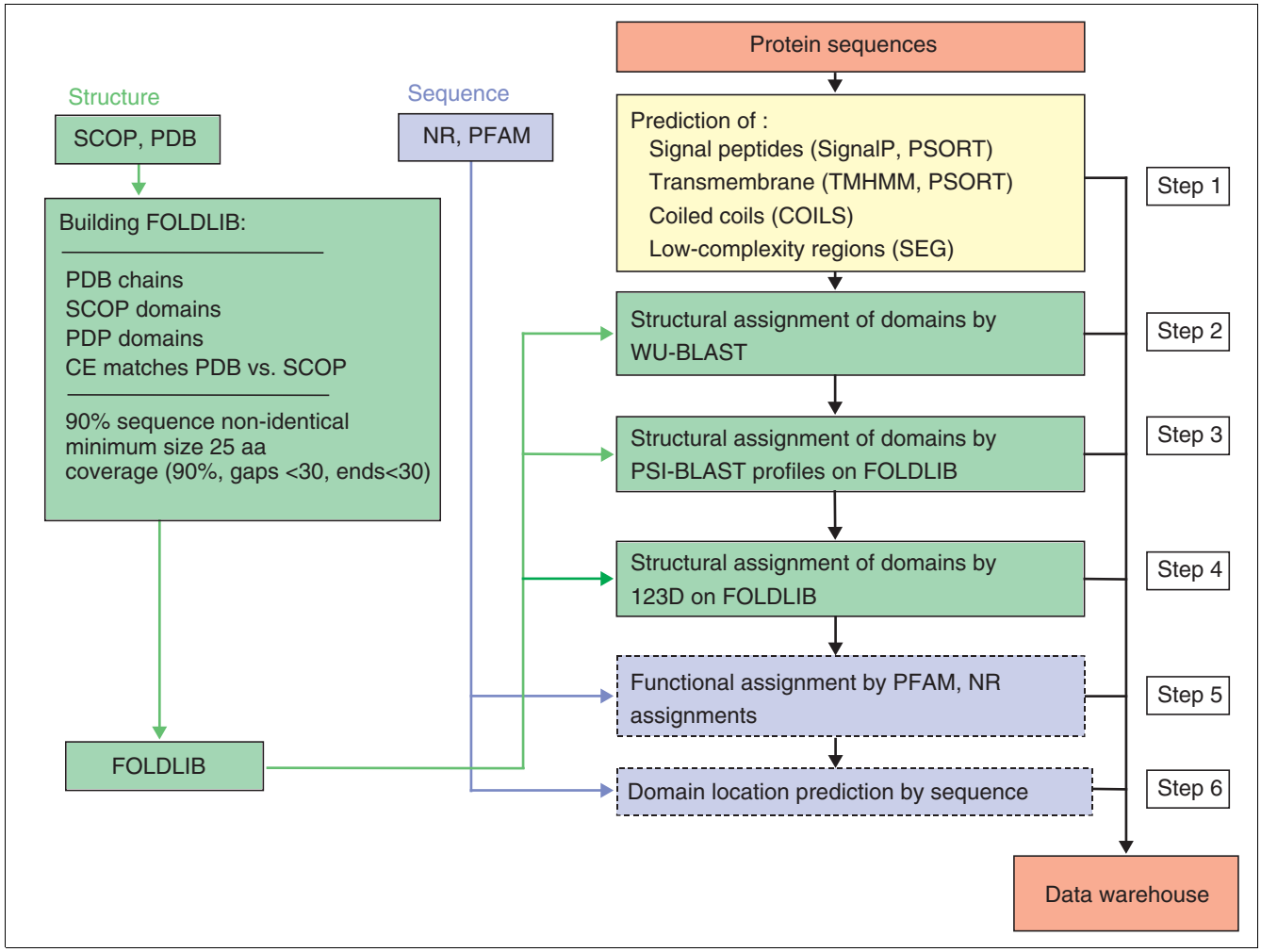

Figure 1
The integrative genome annotation pipeline (iGAP). Processing of initial structural information is shown on the left and processing of initial sequence information on the right. Green shading indicates a processing step involving structure information and blue shading a processing step involving a sequence. Steps boxed with dotted lines indicate partial integration into the benchmarking scheme. See text for further details.

achieved by PAT over other means. Whether these results are meaningful depends on the user's perspective. For one user, a few additional predictions with 90% certainty could be a distraction. To another, they might, in connection with additional experimental evidence, prove valuable. A future challenge to those of us providing such resources is to minimize the pain and maximize the gain for the different types of user. Again quality assurance and user interface design will prove important. While we have made efforts to classify the reliability of our predictions, they are still predictions and should be used, where possible, with associated experimental proof.

With regard to iGAP specifically, we first looked at the overall coverage of the *Arabidopsis* proteome using iGAP (Figure 3). We were able to assign nearly 70% of the *Arabidopsis* proteome to folds which had a reliability index C (90% confidence) or better. This compares to 56% of *Arabidopsis* proteins in the NCBI nonredundant (NR) protein database having an assigned function. While fold assignment does not necessarily translate into functional assignment, it provides a useful indicator.

Second, PAT provides annotations not reported by other databases. Some examples are listed in Table 4. For example, the AP2-domain is a DNA-binding transcription factor that controls flower and seed development [20] in *Arabidopsis*. The structure of the AP2 domain is found in the PDB (1gcc) [21]. Standard BLAST using the 1gcc sequence provides 140 hits at $p < 0.1$ (a very weak threshold). In PAT, there are 143 hits of A or B reliability (> 99% confidence) plus 12 of reliability C (> 90% < 99% confidence). Another putative protein (GI number 15228210, locus id At3g47660) has a previously undetected domain at the amino terminus which resembles the structure of the pleckstrin homology (PH) domain from phospholipase C delta (PDB 1mai) (C prediction). PH

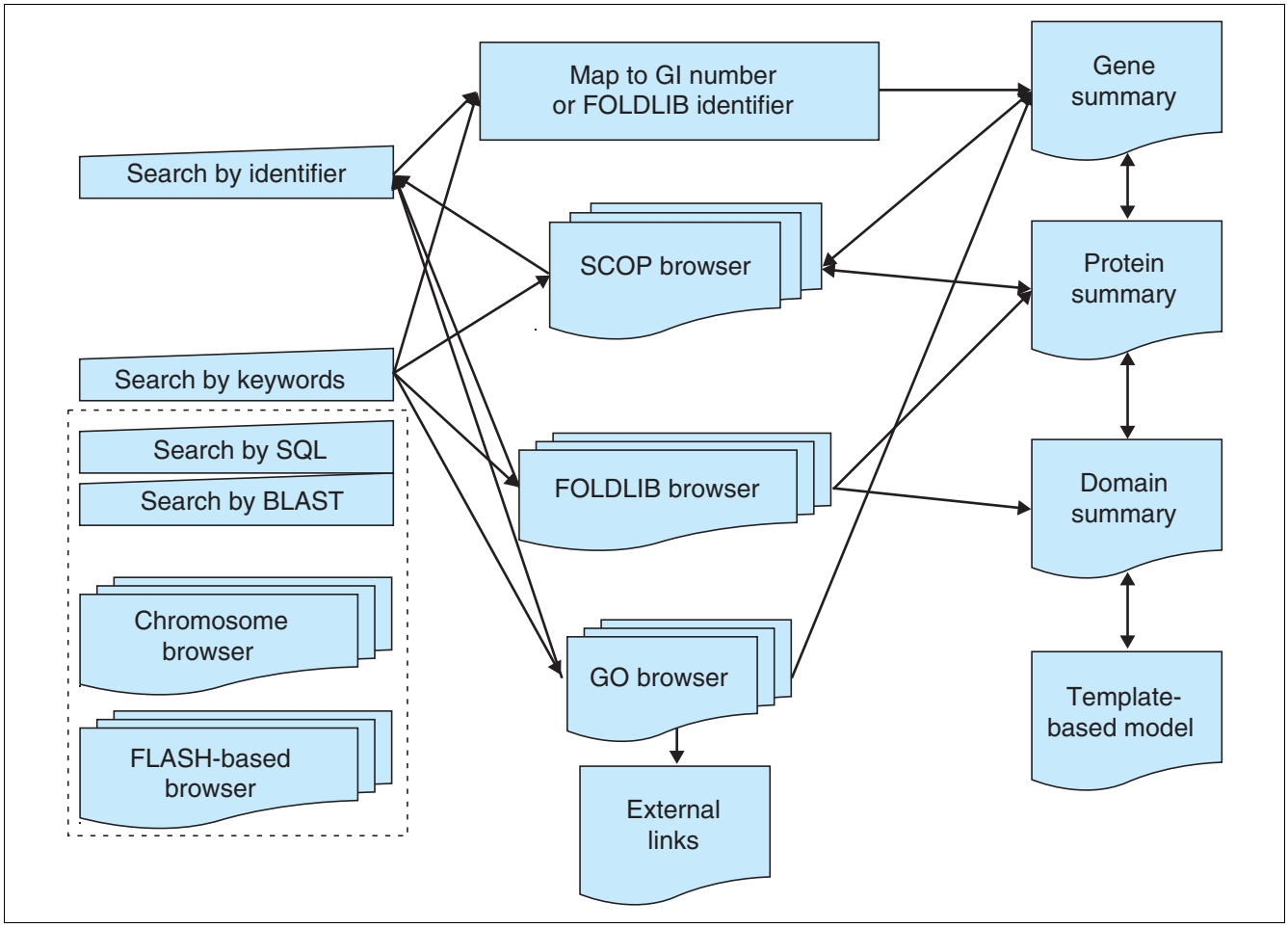

Figure 2
Overview of the user interface. The information stored in the database may be accessed by known identifiers, keywords, browsing classifications (SCOP and FOLDLIB) and by sequence. Identifiers supported include *Arabidopsis* locus id, NCBI gi number, SCOP id, PDB id, FOLDLIB id and PFAM id. Keywords are limited to those available in each original data source.

domains are commonly found in signaling proteins [22]. Additional domains found in this protein (also documented by TAIR as InterPro domains) include FYVE/PHD zinc finger and an RCC1 like domain (a regulator of chromosome condensation), with A and B reliabilities respectively. TAIR also reported a sugar transporter signature for this protein from Prosite. While the exact function of the protein remains to be determined experimentally, the new finding of a putative PH domain could offer clues to its potential mechanism for signaling and intracellular targeting.

Third, we surveyed a set of *Arabidopsis* proteins that have known protein structures (confidence level A, Table 4a). For most of these structures, PAT identifies a number of additional *Arabidopsis* proteins predicted to contain the same domain. For example, the ubiquitin-conjugating enzyme, which is important in protein degradation, identifies 6 unknown proteins out of 12, with 'C' or above confidence, which contain similar domains. In contrast, no additional proteins were found to have TBP-like (TATA binding protein-like) domains.

Recent structures not found in FOLDLIB or SCOP (release 1.55) were examined to see how well they were predicted by iGAP (Table 4b). For PDB structures 1gp4 and 1gp6 (putative leucoanthocyanidin dioxygenase, NCBI NR database 17 October 2001 release), 123D was able to correctly predict the fold to be similar to 1hig (clavaminate synthase-like SCOP superfamily). WU-BLAST only gave a number of low-probability (E reliability) predictions.

Similarly, PDB entry 1e6b (putative glutathione-*S*-transferase, NCBI NR database 17 October 2001) is a protein with an amino-terminal thioredoxin-like domain and a contiguous glutathione-*S*-transferase carboxy-terminal domain. Both WU-BLAST and 123D correctly recognized the template structure 1fw1 (glutathione transferase z/maleylacetoacetate isomerase). Both WU-BLAST and 123D predicted the whole

Table 2

Database feature comparison

Databases	Features	Scope	Level of integration	Learning curve	Drawbacks
Entrez Genome [20]	Domains from CDD (SMART, PFAM) Proteins by NCBI GI number, accession number, Swiss Prot ID, and so on Structure by PDB ID 3D domains from MMDB Domain relatives by CDART Related sequences using BLINK Visualization using Cn3D Public data	All sequences published or voluntarily deposited 1,000+ genomes	High	Easy to high	Complex system Only experimental structural information is available Software interface is not readily available Linkout progress is slow
EBI Proteome Analysis Database [43]	InterPro member databases (SwissProt, PFAM, SMART, TIGRFAM, PRINTS, PROSITE, ProDom, PIR SuperFamily) Families, domains and sites by member databases GO annotation Manual curation and integration Precomputed matches against InterPro entries	Complete proteomes in SwissProt and TrEMBL 110+ proteomes	Medium	Easy to moderate	SRS based query interface free to academia Basic keyword search possible Sequence based classification
MatDB	*Arabidopsis* annotation from PEDANT Free text search Protein categories by structure, function based on SCOP, PIR, InterPro	*Arabidopsis* with limited intergenome comparison	Medium	Easy to moderate	Query response time varies SCOP classification mildly difficult to use
Proteins of *Arabidopsis thaliana* (PAT) database	Domains from SCOP, predicted domains from PDP, and full length PDB chains with less than 90% sequence identity (FOLDLIB) GO annotation Precomputed matches against FOLDLIB Template-based structure models Visualization using QuickPDB, Chime Advanced keyword search Hierarchical browsing based on SCOP Related sequences using WU-BLAST	Currently 87 Expanding to provide coverage for all known proteomes	Medium	Easy to Moderate	Presentation Style Query flexibility implies a higher learning curve
TAIR	GO and other ontology development Sequence and map viewer Domains from InterPro Regulatory motif analysis User annotation	Comprehensive resource devoted to *Arabidopsis*	Medium	Easy to moderate	No structural information
SUPERFAMILY	HMM (SAM) models for SCOP domains Fold recognition Domain architecture visualization	107 genomes	Low to medium	Easy to moderate	Presentation style No update information
Gene 3D	Structural assignment based on CATH domain classification using PSI-BLAST	66 genomes	Low	Easy	Annotation not dynamically linked to CATH No update information

Table 3

Comparison of PAT with other resources

(a) Coverage	PAT	PEDANT/MatDB	TAIR/GO	EBI Proteomes/InterPro
	94% A-E	30.9% PDB	38% ALL	77.3% InterPro
	84% A-D	26.7% SCOP	14% Non-IEA	0.07% PDB
	65% A-C			
	46% A-B			
	38% A			

(b) Specific examples	Target	Other sources	PAT	
			Results	Reliability
	AP2 domain (1gcc)	140 hits by BLAST against NR	155 hits	C (90% certainty) or above
	15239082 (At5g11550.1)	No hits by PSI-BLAST None from TAIR, PEDANT	1EE4	C
	15228210 (At3g47660)	FYVE/PHD zinc finger RCC1 like domain Sugar transporter signature (PROSITE)	FYVE/PHD zinc finger; RCC1 like domain; PH domain	A (99.9% certainty); B (99% certainty); C
	Cytochrome P450	238 (TAIR GO)	249 hits 256 hits	C or above D (50% certainty) or above
	Protein-kinase-like domain	1037 hits (PEDANT/ MatDB) 951 hits (TAIR GO)	1,179 hits	C or above
Alpha/beta hydrolase fold	*Arabidopsis*	194 hits (PEDANT/MatDB, SCOP 3.65)	340 hits 200 hits	C or above A
	Human	69 hits (PEDANT/MatDB, *SCOP* c.69)	1,086 hits 1,18 hits	C or above A

(a) Percent coverage against specific data sources. **(b)** PDB sequence of 1gcc [22] was used to perform a standard BLAST search. The putative protein with gi number 15239082 (At5g11550.1) returns no hits using PSI-BLAST. The putative protein (gi number 15228210, locus id At3g47660) contains a FYVE/PHD zinc finger domain, and an RCC1 like domain (a regulator of chromosome condensation). TAIR also reported a sugar transporter signature for this protein from Prosite search. The term 'cytochrome P450' was used to search TAIR GO annotation (release). This was obtained using the search by keyword query feature, after we've loaded the TAIR GO data into our database. The cytochrome P450 fold in the SCOP hierarchy was used to retrieve the hits from PAT. Actual hits may vary between releases.

protein to be thioredoxin-like with a reliability index of A. However, WU-BLAST made two additional predictions, both correct. The 'pseudo SCOP entry by PAT' is a novel domain parsed by PDP, which at the time was not in SCOP release 1.55. (It is classified as a separate domain in SCOP 1.59.) This was recognized by WU-BLAST. Additionally, WU-BLAST also recognized the amino-terminal thioredoxin-like domain with correct boundaries.

Finally, the SCOP classification of protein structures by fold (Figure 4a) and by family (Figure 4b) provides a convenient way to catalog the relative occurrences of structures in *A. thaliana*. With respect to folds, the membrane all-alpha fold, alpha-alpha superhelix and protein kinase-like (PK-like) fold ranked highest. The TIM barrel and Rossman folds, and seven-bladed beta-propeller folds are also among the top folds. PK-like proteins have the second highest occurrence at the superfamily level (data not shown). Not surprisingly, serine/threonine kinases and tyrosine kinases are among the most abundant families.

Conclusions

The PAT database was initially developed as a joint development of academia and industry to serve the *Arabidopsis* and plant proteomics community through the provision of structure and functional assignment to all identified proteins in the *Arabidopsis* genome. The underlying technology, specifically iGAP and the associated reliability criteria, is well suited for application to other proteomes and this processing is ongoing to provide a comparative proteomics resource. With more of a focus on comparative proteomics, the resource is being expanded in an effort we refer to as the Encyclopedia of Life (EOL). Details on EOL can be found at [23].

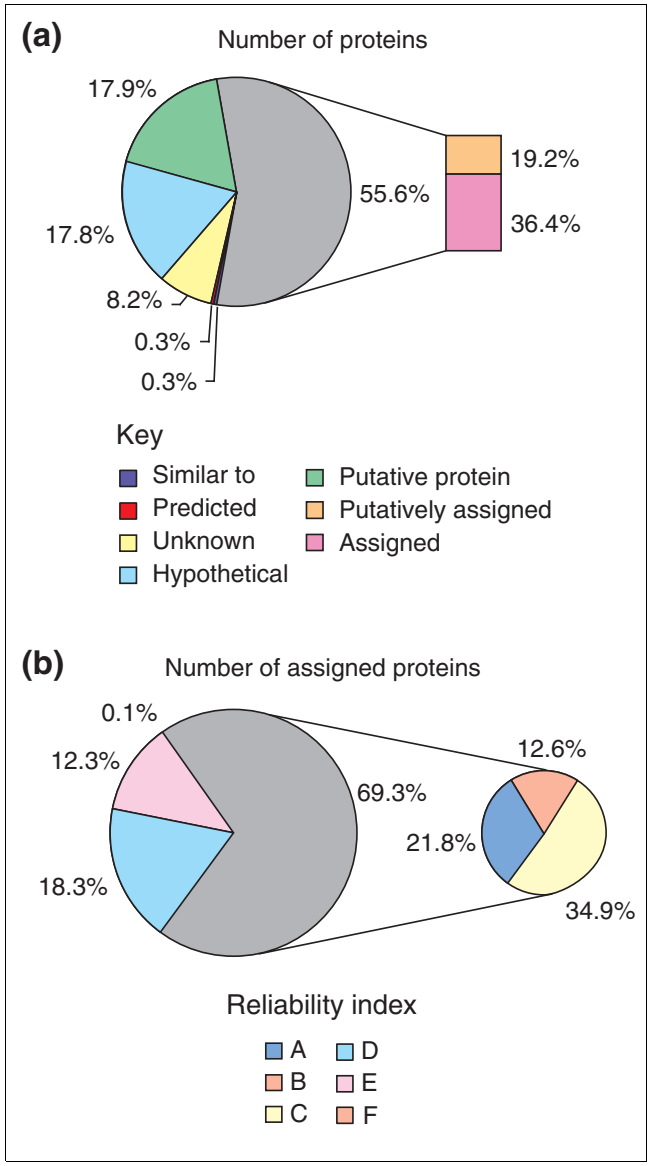

Figure 3
Classes of *Arabidopsis* proteome annotation. **(a)** The functional annotation on *Arabidopsis* proteins provided by the NCBI NR database. In this database, 36.4% of *Arabidopsis* proteins are reliably assigned on the basis of experimental evidence; 55.6% are annotated when automated annotation is included. This data is based on the 17 October 2001 release of NR. **(b)** Structural annotation provided by PAT. PAT has 69.3% coverage with a C reliability or better.

Materials and methods

The iGAP components are shown in Figure 1, which illustrates how primary protein sequence and structure data are processed by the system. Details are given below.

Software and availability

The software components of iGAP have been tested on Redhat Linux 7.2, Sun Solaris 5.8 and the IBM AIX operating systems. It is currently ported to the Teragrid platform [24] for high-performance distributed computing. Access is via an Apache web server (1.3.25) and an Oracle 9.2.0 database at the San Diego Supercomputer Center where high uptime is maintained. A new interface based on Java 2 Enterprise Edition (J2EE) and Struts framework is under development.

The iGAP software components developed at the University of California San Diego (UCSD) are available free for academic use by contacting the authors as part of the University of California Copyright Agreement. For-profit organizations need to contact the UCSD Technology Transfer Office. Separate licenses may be required for non-UCSD components. The key components and steps are described below, with additional details available from the Web [25].

FOLDLIB

SCOP domain sequences filtered at 90% identity [26] are downloaded from the Astral database [27]. PDB chains are clustered at 90% identity and parsed with PDP [15] to provide additional domains, including those not yet assigned by SCOP. SCOP lags behind the PDB in terms of structures processed. The sequences from SCOP, PDB, and PDP are then clustered at 90% identity to define the final structure-template library. Profile libraries for these templates are generated for use by 123D using PSI-BLAST with a default E-value of 1e-6 and three iterations.

The pipeline

The first step of the pipeline uses a set of filter programs to determine the low-complexity regions as well as transmembrane regions, signal-peptide sequences, and coiled coils in a particular proteome. The programs used include SEG [28] for low-complexity region, COILS [29] for coiled coils, TMHMM [30] for transmembrane region, PSORT [31] for subcellular location and signalP [32] for signal peptides.

The second step determines sequence similarity hits by pairwise sequence comparison using WU-BLAST (W. Gish, personal communication). WU-BLAST is used because it is fast and performed best in our benchmark studies. The default E-value used is 1e-5. The third step generates PSI-BLAST profiles for each input protein sequence against the FOLDLIB sequences. The default H-value used is 1e-6 and three iterations for profile generation. In the fourth step, the program 123D is used to provide additional mapping to FOLDLIB using fold recognition [17]. 123D has been used successfully in CASP [33] competitions.

Reliability index

The reliability of a prediction is calculated on the basis of a novel benchmarking procedure against SCOP and will be described elsewhere. The index is expressed as percent certainty that a particular prediction is correct: A = 99.9% certainty, B = 99% certainty, C = 90% certainty, D = 50% certainty, and E = 10% certainty.

Table 4

Sampling of known Arabidopsis protein structures in PAT

(a) PDB structures from *Arabidopsis* mapped to FOLDLIB entries	PDB ID	SCOP family	SCOP superfamily	GI number	Name	Domain found	Reliability	Number of unknown or putative proteins with similar domain : total number*
	1dj2	Nitrogenase iron protein-like	P-loop containing nucleotide triphosphate hydrolases	15230358	Adenylosuccinate synthetase	1dj2 (48-490)	A	1:2
	1dcf	The receiver domain of the ethylene receptor	CheY-like	15219629	The receiver domain of the ethylene receptor	1dcf (605-736)	A	19:33
	1jh7	Cyclic nucleotide phospho-diesterase	Cyclic nucleotide phospho-diesterase	15234068	Putative protein	1fsi (1-181)	A	2:2
	2aak	Ubiquitin conjugating enzyme	Ubiquitin conjugating enzyme	15223746	Ubiquitin conjugating enzyme	1a3s (1-151)	A	6:12
	1vok	TATA-box binding protein (TBP), carboxy-terminal domain	TATA-box binding protein-like	15231241	TATA sequence-binding protein 1	1ais (12-198)	A	0:2
	3nul	Profilin (actin-binding protein)	Profilin (actin-binding protein)	15224838	Profilin 1	3nul 2-131)	A	0:4
	1ibj	Cystathionine synthase-like	PLP-dependent transferases	15230203	Cystathionine beta-lyase precursor	1ibj (1-464)	A	41:54
(b) PDB structures not found in FOLDLIB	PDB ID	SCOP family	SCOP superfamily	GI number	Name	Domain found	Reliability	Method
	1gp4,6	Penicillin synthase-like	Clavaminate synthase-like	15235853	Putative leucoantho-cyanidin dioxygenase	1hjg (43-350)	A	123D
	1e6b (88-220)	Glutathione *S*-transferases, carboxy-terminal domain	Pseudo SCOP entry by PAT (glutathione *S*-transferases, carboxy-terminal domain)	15226952	Putative glutathione *S*-transferase	1fw1 (89-193)	A	WU-BLAST
		Thioredoxin-like (glutathione *S*-transferases, carboxy-terminal domain)				1fw1 [1-218]	A	123D
						1fw1 [11-215]	A	WU-BLAST
	1e6b (8-87)	Thioredoxin-like				1fw1 (11-89)	A	WU-BLAST

(a) The known *Arabidopsis* PDB ids are obtained from NCBI pdbaa FASTA file (9/1/02 release). Each PDB id is used as a query using the PAT id search field. The 'Domain found' column lists some of the domains found in the protein. Use the GI number to search the PAT web site to see all possible domain assignments. If there are multiple domain boundaries specified, only the longest possible domain boundary is listed. *Non-NR entries were also excluded in the statistics collected in the last column of the table. Only predictions with higher than C reliability (90% certainty) are included. The non-NR entries (contributed by Ceres, Inc) were absent from NR of NCBI at the time of analysis. 1gp4, 1gp6, and 1e6b were not in SCOP release 1.55 or the FOLDLIB in this study (see Table 1b). 1j6y was an NMR structure and was excluded. **(b)** The sequences of the three structures not in the FOLDLIB were analyzed as unknown proteins. The assignment by SCOP release 1.59 is enclosed in parenthesis. In the case of 1e6b, two distinct domains are classified by SCOP 1.59. The two regions are listed after the PDB id. In the case of 1gp4 or 1gp6, only 123D produced an A prediction correctly. In the case of 1e6b, the template is predicted correctly by both 123D and WUBLAST, but WUBLAST produced multiple domains, two of which coincides with SCOP release 1.59 assignment.

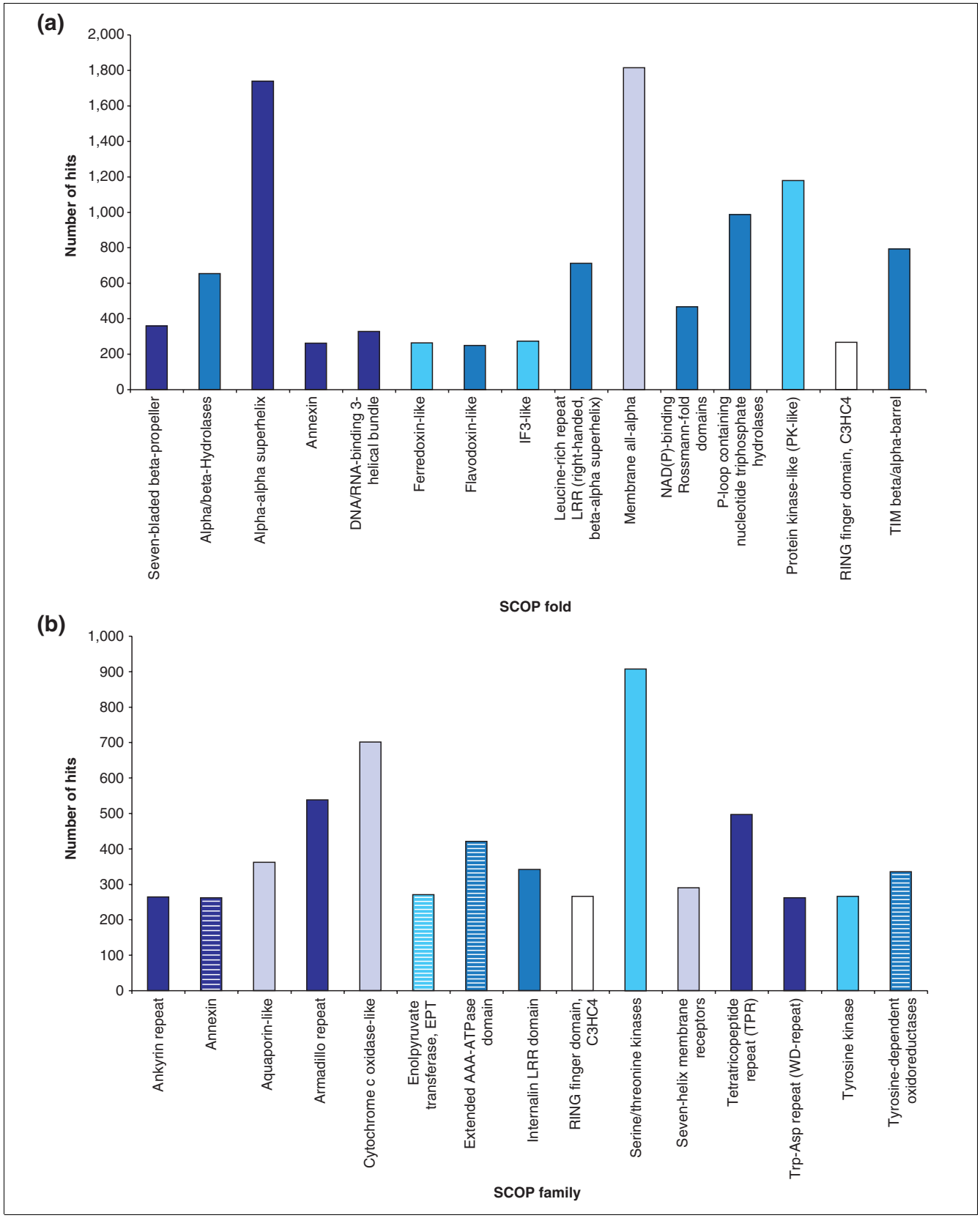

Figure 4 *(see legend on next page)*

Figure 4 *(see previous page)*
SCOP classifications for the *Arabidopsis thaliana* proteome. **(a)** Occurrences of SCOP folds. Folds belonging to the same SCOP class are shaded the same color. **(b)** Occurrences of SCOP families. Families belonging to the same fold are shaded the same color. Families belonging to the same fold but to different superfamilies are indicated by striped bars. The top 15 folds and families are shown. Data are based on SCOP release 1.59.

Database and user interface

Data provided by iGAP are stored in an Oracle 9i (release 2) relational database system. The database is connected to the web using Apache mod_perl and the Perl DBI. External data sources include SCOP, NR, PFAM, NCBI taxonomy, LocusLink [34], SwissProt [35] and InterPro [36].

Chromosomal position information for the *Arabidopsis* data were obtained from the TIGR *Arabidopsis thaliana* database [37]. The physical and chemical properties are calculated using the EMBOSS pepstats program [38]. The Gene Ontology assignment for *Arabidopsis* was obtained from The *Arabidopsis* Information Resource (TAIR) [39]. We have also developed our own methodology for assigning additional GO terms with a measure of likelihood (W Krebs and P.E.B., unpublished work) beyond those assigned by SwissProt.

By default, only those predictions with a reliability index of C or above are shown. The reliability index for all queries may be changed using a pull down menu. The key characteristics of the Web interface that we have developed include the following (Figure 2).

SCOP browser

The use of SCOP classifications provides a hierarchical view of the data from a structure perspective. For example, the user may start with the all-alpha class and drill down through fold, superfamily, family, and domain level. Alternatively, the structure classification can be searched for terms such as "Rossman fold" present in SCOP annotation.

FOLDLIB browser

The classification of protein folds in the fold library can be browsed. Alternatively, it can be searched by PDB id or sequence.

Search by identifier

The database may be searched using identifiers from a number of existing databases such as SCOP, PFAM (ID or Accession Number), NCBI (GI number), PDB identifier, Locus identifier, Gene Ontology (GO) term [40], or FOLDLIB identifier.

Search by keywords

Descriptions from NR, PFAM, PDB, FOLDLIB, SCOP and GO are parsed and indexed. The text index supports complex searches and wild card searches. No attempt is made to reconcile nomenclature differences introduced by each individual data source.

Domain summary

This provides preliminary information on a particular domain, identified by its FOLDLIB id. The protein domain sequence is displayed and its structure may be viewed using a Chime (MDL, San Leandro, CA) plug-in [41]. All sequences which contain the same domain are displayed. For each sequence, a link provides the specific target-template alignment and a graphic representation of the domain architecture. It also links to the template based models described below.

Gene summary

This provides preliminary information on all the domains located within a particular gene including domain boundary information. Each domain may subsequently be interrogated with the SCOP browser to provide superfamily, family and fold level information. The protein summary page provides comprehensive information about the protein besides domain assignment.

Template-based models

From the template target alignment, 3D coordinates from the FOLDLIB template are used to construct a C-alpha only PDB format file using the sequence of the target protein. The resulting PDB file may then be visualized using QuickPDB, a Java applet developed by I.N.S. and P.E.B. (unpublished), or with other popular 3D viewers such as the Chime viewer plugin.

Availability and update

The data are available from the Web [25]. Information may be downloaded in text or XML format and imported into an Excel spreadsheet, MySQL database or other applications. For advanced users, the data may be retrieved using SQL from the Web interface. A database schema is available on the SQL search page as an aid in SQL query formulation.

A workflow management system is under development to automate the processing and update of proteomes. All external data are updated when a major release of NR becomes available. NR database is downloaded from NCBI. Sequences from other sequencing centers are clustered at 100% identity using cd-hit [42]. Subsequent updates are performed monthly using the NCBI NR Month database. The unique sequences are sorted according to taxonomy using the NCBI gi_taxonomy mapping table. Only sequences that are new or changed (crc64 checksum) are submitted to a continuous update process. The release date for each source database used is given on the home page. The *Arabidopsis* proteome

(27,242 total and 27,089 unique sequences, 7 September 2002 release) may be computed in approximately 50,000 computer hours.

Acknowledgements

This work is supported by the National Partnership for Advanced Computational Infrastructure (NPACI) funded by the National Science Foundation (NSF) grant ASC 9619020 and the National Institutes of Health (NIH) grant GM63208-01A1S1. The authors wish to thank the many biologists who provided feedback to the development of the database and interface, the authors of the external software components, and Robert Byrnes for reviewing the manuscript.

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
2. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
3. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002.** *Nucleic Acids Res* 2002, **30:**235-238.
4. Pietrokovski S, Henikoff JG, Henikoff S: **The Blocks database - a system for protein classification.** *Nucleic Acids Res* 1996, **24:**197-200.
5. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
6. Aravind L, Dixit VM, Koonin EV: **Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons.** *Science* 2001, **291:**1279-1284.
7. Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ, Ledley RS, Lewis KC, Mewes HW, Orcutt BC, *et al.*: **The Protein Information Resource: an integrated public resource of functional annotation of proteins.** *Nucleic Acids Res* 2002, **30:**35-37.
8. Hoersch S, Leroy C, Brown NP, Andrade MA, Sander C: **The GeneQuiz web server: protein functional analysis through the Web.** *Trends Biochem Sci* 2000, **25:**33-35.
9. Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A, Mewes HW: **Functional and structural genomics using PEDANT.** *Bioinformatics* 2001, **17:**44-57.
10. Gough J, Chothia C: **SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments.** *Nucleic Acids Res* 2002, **30:**268-272.
11. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource.** *Nucleic Acids Res* 2002, **30:**242-244.
12. **3D-Genomics** [http://www.sbg.bio.ic.ac.uk/3dgenomics]
13. Kelley LA, MacCallum RM, Sternberg MJ: **Enhanced genome annotation using structural profiles in the program 3D-PSSM.** *J Mol Biol* 2000, **299:**499-520.
14. Buchan DW, Shepherd AJ, Lee D, Pearl FM, Rison SC, Thornton JM, Orengo CA: **Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database.** *Genome Res* 2002, **12:**503-514.
15. Alexandrov N, Shindyalov I: **PDP: protein domain parser.** *Bioinformatics* 2003, **19:**429-430.
16. Shindyalov IN, Bourne PE: **A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm.** *Nucleic Acids Res* 2001, **29:**228-229.
17. Alexandrov NN, Fischer D: **Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures.** *Proteins* 1996, **25:**354-365.
18. Alexandrov NN, Luethy R: **Alignment algorithm for homology modeling and threading.** *Protein Sci* 1998, **7:**254-258.
19. **NCBI Genomic Biology** [http://www.ncbi.nih.gov/Genomes]
20. Okamuro JK, Caster B, Villarroel R, Van Montagu M, Jofuku KD: **The AP2 domain of APETALA2 defines a large new family of DNA binding proteins in *Arabidopsis*.** *Proc Natl Acad Sci USA* 1997, **94:**7076-7081.
21. Allen MD, Yamasaki K, Ohme-Takagi M, Tateno M, Suzuki M: **A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA.** *EMBO J* 1998, **17:**5484-5496.
22. Mayer BJ, Ren R, Clark KL, Baltimore D: **A putative modular domain present in diverse signaling proteins.** *Cell* 1993, **73:**629-630.
23. **The Encyclopedia of Life Project** [http://eol.sdsc.edu]
24. **TeraGrid** [http://www.teragrid.org]
25. **Proteins of *Arabidopsis thaliana* (PAT) Database** [http://pat.sdsc.edu]
26. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2002: refinements accommodate structural genomics.** *Nucleic Acids Res* 2002, **30:**264-267.
27. Chandonia JM, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE: **ASTRAL compendium enhancements.** *Nucleic Acids Res* 2002, **30:**260-263.
28. Wootton JC, Federhen S: **Analysis of compositionally biased regions in sequence databases.** *Methods Enzymol* 1996, **266:**554-571.
29. Lupas A, Van Dyke M, Stock J: **Predicting coiled coils from protein sequences.** *Science* 1991, **252:**1162-1164.
30. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6:**175-182.
31. Nakai K, Horton P: **PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization.** *Trends Biochem Sci* 1999, **24:**34-36.
32. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Int J Neural Syst* 1997, **8:**581-599.
33. Moult J, Fidelis K, Zemla A, Hubbard T: **Critical assessment of methods of protein structure prediction (CASP): round IV.** *Proteins* 2001, **Suppl 5:**2-7.
34. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
35. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its supplement TrEMBL.** *Nucleic Acids Res* 1997, **25:**31-36.
36. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, *et al.*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Res* 2001, **29:**37-40.
37. **The Institute for Genomic Research** [http://www.tigr.org]
38. **EMBOSS: The European Molecular Biology Open Software Suite** [http://www.hgmp.mrc.ac.uk/Software/EMBOSS/]
39. **TAIR: The Arabidopsis Information Resource** [http://www.arabidopsis.org]
40. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
41. **The MDL Chime Site** [http://www.mdl.com/chime]
42. Li W, Jaroszewski L, Godzik A: **Clustering of highly homologous sequences to reduce the size of large protein databases.** *Bioinformatics* 2001, **17:**282-283.
43. **EBI Proteome Analysis Database** [http://www.ebi.ac.uk/proteome]