W444–W450 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm364

ProtSweep, 2Dsweep and DomainSweep: protein analysis suite at DKFZ

C. del Val[1,2,*], P. Ernst[1], M. Falkenhahn[1], C. Fladerer[1], K. H. Glatting[1], S. Suhai[1] and A. Hotz-Wagenblatt[1]

[1]DKFZ, German Cancer Research Center, Division of Molecular Biophysics, Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany and [2]Department of Computer Science and Artificial Intelligence, ETSII University of Granada, C/Daniel Saucedo Aranda s/n 18071, Granada, Spain

Received January 31, 2007; Revised April 4, 2007; Accepted April 25, 2007

ABSTRACT

The wealth of transcript information that has been made publicly available in recent years has led to large pools of individual web sites offering access to bioinformatics software. However, finding out which services exist, what they can or cannot do, how to use them and how to feed results from one service to the next one in the right format can be very time and resource consuming, especially for non-experts.

Automating this task, we present a suite of protein annotation pipelines (tasks) developed at the German Cancer Research Centre (DKFZ) oriented to protein annotation by homology (ProtSweep), by domain analysis (DomainSweep), and by secondary structure elements (2Dsweep). The aim of these tasks is to perform an exhaustive structural and functional analysis employing a wide variety of methods in combination with the most updated public databases. The three servers are available for academic users at the HUSAR open server http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar/

INTRODUCTION

As more and more genomes are automatically sequenced, comprehensive protein annotation is a needed step after gene identification. Even in good annotated genomes (human, mouse) about 30% of all proteins are not functionally identified (1–3), and thus often a similarity search will not be sufficient. Here, we present a suite of protein tasks, ProtSweep, DomainSweep and 2Dsweep, which perform analysis from sequence similarity to small domains and structural elements. This includes similarity searches against protein sequence databases and specialized motif collections, prediction of secondary structural elements, attributing each sequence to known superfamilies, protein localization prediction, physicochemical protein characteristics and domain functional assignation. Our strategy for assigning relevant functional roles is based on the joint use of both global (homology similarity) and local (domain and motif) sequence similarities (4). The three servers are available for academic users at the HUSAR open server http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar/

WEB INTERFACE

The input for all the three servers is a protein sequence. Several query sequences can be uploaded by the usual 'copy & paste' procedure into the input box using FASTA format. If more than one sequence is to be queried, a multiple FASTA file can be used. The query starts by clicking on the 'submit' button. Then the user will be redirected to an application page, and the run" button can start the task. Additionally, there is a link to an online help, indicated with a '?', with the following topics: *short description*, *programs employed*, *algorithm*, *output*, *additional options* and *acknowledgments*. Results can be received by selecting the tab 'Go to results page'. The results are provided as HTML for visual inspection or can be downloaded as XML for storage in private databases. In case of error when clicking on the application name in the Results Manager page, a log-file is displayed where more human readable error messages can be found.

Databases

Standard protein databases used by the tasks like Uniprot/SwissProt, Uniprot/TREMBL and RefSeq are automatically updated whenever new versions become

*To whom correspondence should be addressed. Tel: +49-6221-422349; Fax: +49-6221-439633; Email: c.delval@dkfz.de, delval@decsai.ugr.as
Present address:
M. Falkenhann, Sanofi-Aventis Deutschland GmbH, Germany
C. Fladerer, Proteom Centrum Tübingen, Germany

Table 1. Databases used in the different pipelines, and the programs and parameters used to search them

Server	Databases	Links	Program	Parameters
ProtSweep	Uniprot/Swissprot	http://www.ebi.ac.uk/swissprot	BlastP	-NOFILTER -EXP = 10.0
	Uniprot/Trembl	http://www.ncbi.nlm.nih.gov/RefSeq/	BlastP	-NOFILTER -EXP = 10.0
	Refseq	ftp://ftp.ebi.ac.uk/pub/databases/trembl/sptrembl/	BlastP	-NOFILTER -EXP = 10.0
	Ensembl	http://www.ensembl.org/	BlastP	-NOFILTER -EXP = 10.0
2DSweep	DSSP	http://www.sander.ebi.ac.uk/dssp/	BlastP	-EXP = 0.001 -EXTension = 10 -NOGAPPEDalign
	Nrpep (non-redundant NCBI protein database)	ftp://ftp.ncbi.nih.gov/blast/db/FASTA	PsiBlast	-b20000 -a5 -j2 -e0.001
	Uniprot/Swissprot +Uniprot/Trembl +updates	http://www.expasy.ch/sprot/sprot-top.html http://www.ebi.ac.uk/swissprot	MSFGenerator	-EXP = 0.001 -OVERLAP = 75 -CUSTOMRANGE = 5 CUSTOMPERCENTAGE = 80,60,50,45,40,35
DomainSweep	Prosite	ftp://ftp.expasy.ch/databases/prosite	Motifs & Pfscan	Default
	Pfamhmm	http://www.sanger.ac.uk/Software/Pfam/	HMMscan	-lib = pfam.hmm -d
	Prints	ftp.bioinf.man.ac.uk	HMMscan	-lib = prints.hmm -d
	Smart	http://smart.embl-heidelberg.de/	HMMscan	-lib = smart.hmm -d
	Tigrfams	http://www.tigr.org/TIGRFAMs/index.shtml/	HMMscan	-lib = tigrfams.hmm -d
	SCOP	http://scop.mrc-lmb.cam.ac.uk/scop	SCOPscan	Default
	Blocks	ftp://ftp.ncbi.nih.gov/repository/blocks/unix	Blockssearcher	-cutoff = 0.01 –d
	Interpro	http://www.ebi.ac.uk/interpro/	SRS queries	
	Prodom	http://www.toulouse.inra.fr/prodom.html	Prodomblast	Default

available. Concerning EnsEMBL, the situation is more complex. Due to possible inconsistencies between the different EnsEMBL API versions, which are used in the tasks, it is not possible to automatically update the EnsEMBL data and this needs to be done by hand. The different databases and the way they were used in these pipelines are described in Table 1.

APPLICATION PIPELINES

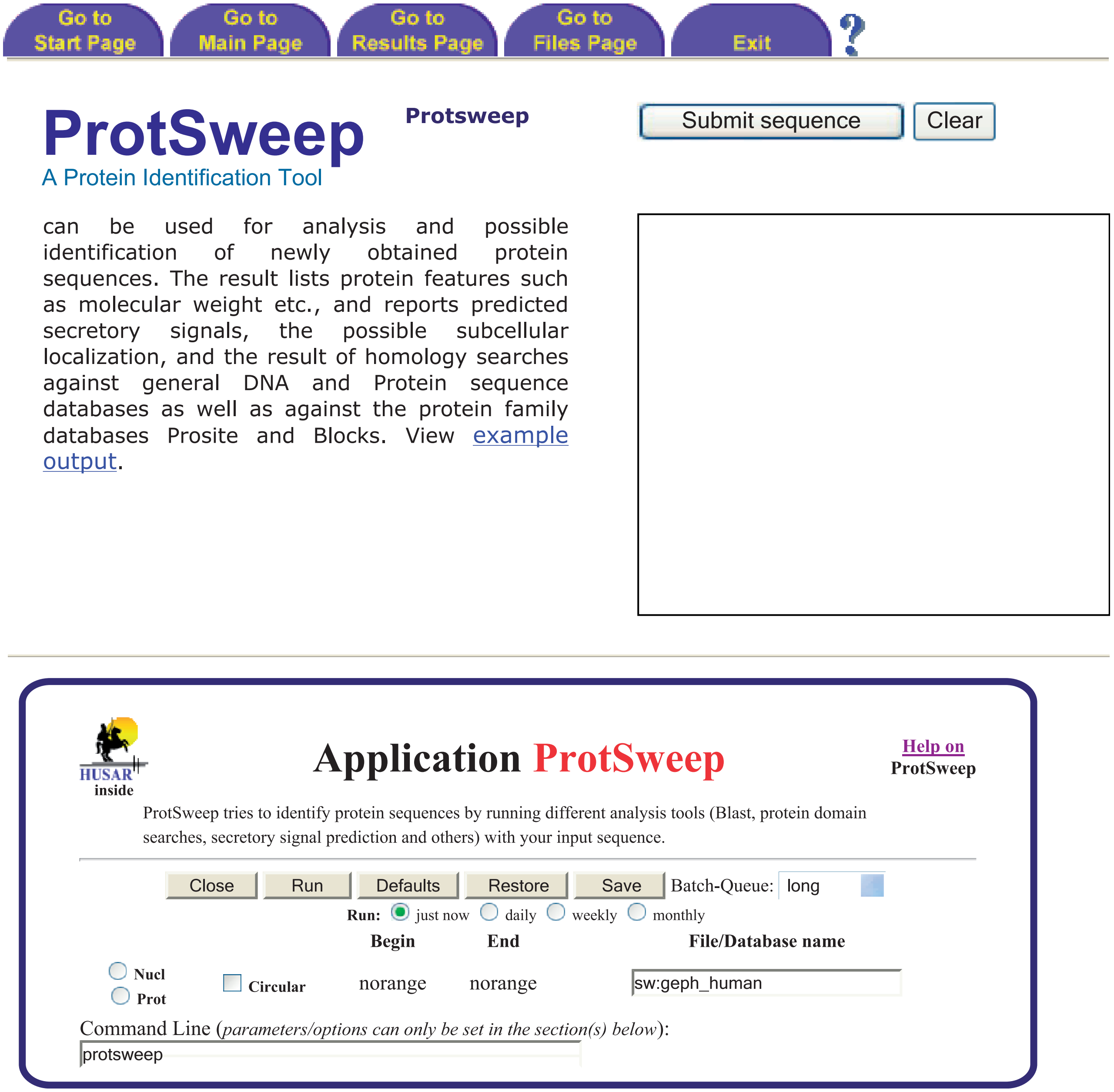
ProtSweep is an approach to the functional characterization of unknown proteins based on a cascade of similarity searches. It is well known that protein databases do not completely overlap and differ in their annotation quality (5). This task takes into account the significant differences among databases (Supplementary Table 1) to improve the quality of the protein characterization. It selects the order in which the databases have to be searched and combines the annotation found depending on the results.

Protsweep classifies proteins into the following categories: *identical*, *homolog*, *similar, weakly similar* and *putative* proteins. The query protein starts the BLAST (6) cascade against Swissprot (7) first (Figure 2). We do take into account three parameters to classify the BLAST hits: (i) percentage of identity, (ii) '*qpercent*' and (iii) '*spercent*'. The two last parameters are related to the length of the total alignment, being '*qpercent*' the percentage of the query sequence length covered in the alignment with the database hit and '*spercent*' the percentage of the hit (subject) sequence length covered by the alignment (Figure 2). Depending on the classification of the BLAST hits according to these parameters and the hit protein annotation, three different approaches will be followed.

If the hit has 100% '*qpercent*' and '*spercent*' and more than 98% identity, it is considered an identical protein and the Swissprot ID will be searched in Ensembl (8). If it is successful, all information from both databases will be combined (Supplementary Table 2) and stored in the XML output. If the ID cannot be found in Ensembl then a BLAST search is performed with the query protein against Ensembl. The best Ensembl hit is selected and compared against the Swissprot hit using the Smith–Waterman algorithm implemented in Water (EMBOSS) (9). If the identity between sequences is greater than 98%, then the information from both sources and the BLAST alignment will be added to the final output, if the identity is less, only Swissprot annotation and the alignments will be added to the XML. If the '*qpercent*' and '*spercent*' is between 80% and 98% and the identity is between 85% and 98%, the hit is classified as *homologous* and follows the same strategy with Ensembl as already described (Figure 2). In case, the identity is between 20% and 85% and '*qpercent*' and '*spercent*' are greater than 85%, then the BLAST cascade continues with SpTrembl and RefseqProt. In the case that no *identical* or *homologous* hits can be found in any of the databases, the best similar hit among the three databases is selected and classified as *similar, weakly similar* or *putative* (Figure 2).

Depending on the classification, the task displays different kinds of information. If the protein is characterized, information concerning the coding gene, about the splicing variants and orthologous genes is also provided. Depending on the degree of homology, protein function, transcript of origin, genomic localization, and GO annotation or partial similarities will also be shown. Proteins annotated as 'hypothetical' are further analysed. Hypothetical proteins will only be presented in the result when no other information about identical or homologous proteins can be found in any of the databases (Supplementary Figure 1).

The web output of ProtSweep (Supplementary Figure 1) is divided in five sections: (i) General Information,



ProtSweep

A Protein Identification Tool

Protsweep

can be used for analysis and possible identification of newly obtained protein sequences. The result lists protein features such as molecular weight etc., and reports predicted secretory signals, the possible subcellular localization, and the result of homology searches against general DNA and Protein sequence databases as well as against the protein family databases Prosite and Blocks. View example output.

Application ProtSweep

Help on ProtSweep

ProtSweep tries to identify protein sequences by running different analysis tools (Blast, protein domain searches, secretory signal prediction and others) with your input sequence.

Batch-Queue: long

Run: ● just now ○ daily ○ weekly ○ monthly

		Begin	End	File/Database name
○ Nucl ○ Prot	□ Circular	norange	norange	sw:geph_human

Command Line (*parameters/options can only be set in the section(s) below*):

protsweep

Figure 1. Outlook of the application pages for the pipeline analyses.

(ii) Identified Protein and Transcripts, (iii) Features and Functions, (iv) Genomic Localisation and (v) Homology to Other Organisms/Genes. The information provided in each of these sections is provided in Figure 2 and Supplementary Table 2. The user has immediate access to all complete application outputs and database entries via hyperlinks. At the bottom of the HTML output there is a link to the explanatory legend as well as to the XML output containing all the generated information.

DomainSweep identifies the domain architecture within a protein sequence and therefore aids in finding correct functional assignments for uncharacterized protein sequences (Figure 3). It employs different database search methods to scan a number of protein/domain family databases. Among these models, in increasing complexity, are: PRODOM (10), automatically generated protein family consensus sequences, PROSITE (11) regular-expression patterns, BLOCKS (12), ungapped position-specific scoring matrices of sequence segments, PRINTS (13) sequence motifs, PROSITE profiles (7), gapped position-specific scoring matrices and Hidden Markov Models like PFAM (14), SMART (15), TIGRFAMS (16) and SCOP (17). Each database covers a slightly different, but overlapping set of protein families/domains. Each model has its own diagnostic strengths and weaknesses and for each of these protein/domain family databases used we have established different thresholds. For example, in the case of the database PFAM-A, we compare the input sequence against the Hidden Markov model profile of each PFAM protein family.

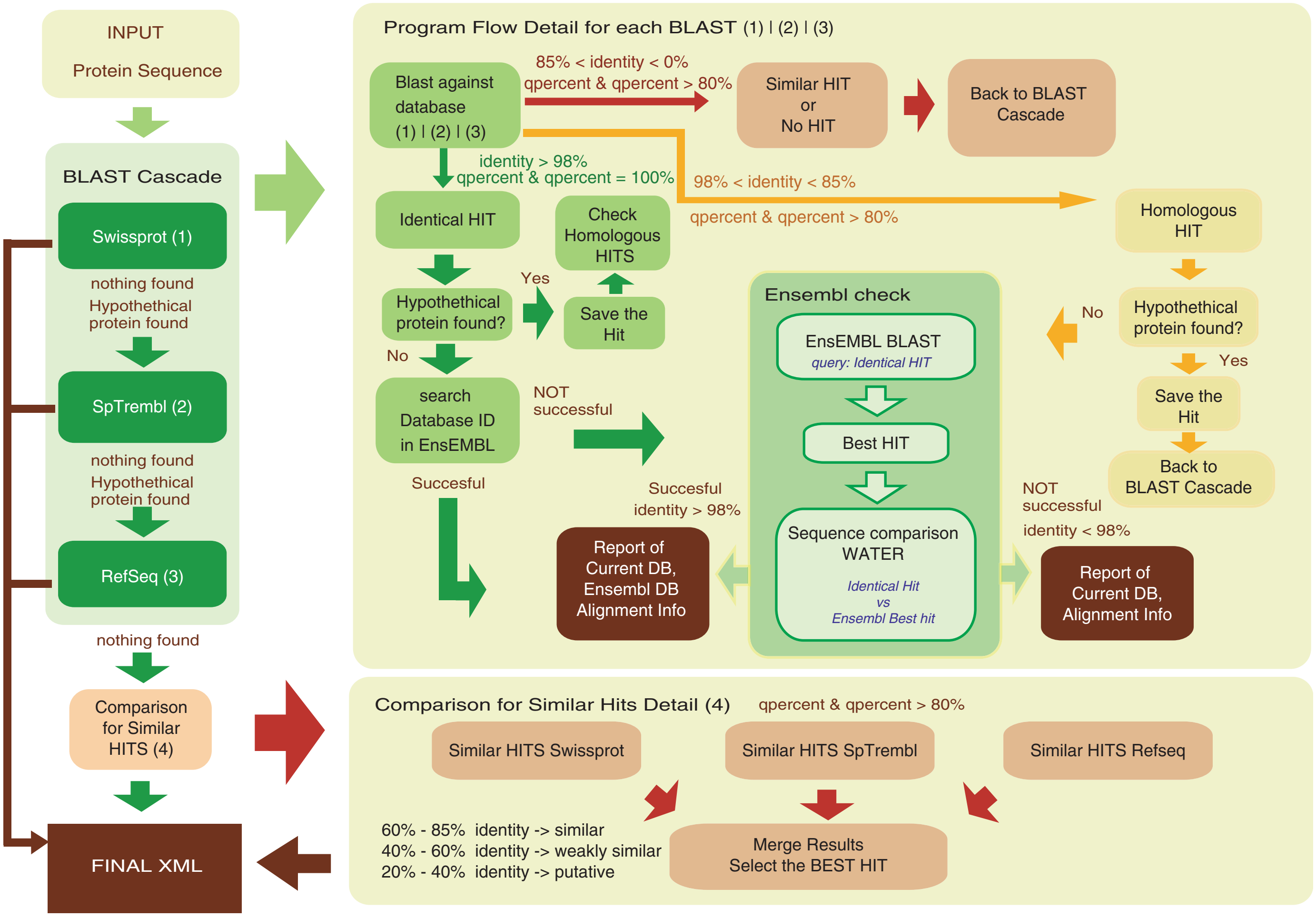

INPUT
Protein Sequence
BLAST Cascade
Swissprot (1)
nothing found
Hypothethical protein found
SpTrembl (2)
nothing found
Hypothethical protein found
RefSeq (3)
nothing found
Comparison for Similar HITS (4)
FINAL XML
Program Flow Detail for each BLAST (1) | (2) | (3)
Blast against database (1) | (2) | (3)
85% < identity < 0%
qpercent & qpercent > 80%
Similar HIT or No HIT
Back to BLAST Cascade
identity > 98%
qpercent & qpercent = 100%
98% < identity < 85%
qpercent & qpercent > 80%
Homologous HIT
Identical HIT
Check Homologous HITS
Yes
Hypothethical protein found?
Save the Hit
No
search Database ID in EnsEMBL
NOT successful
Succesful
Ensembl check
EnsEMBL BLAST
query: Identical HIT
Best HIT
Sequence comparison WATER
Identical Hit vs Ensembl Best hit
Succesful
identity > 98%
Report of Current DB, Ensembl DB Alignment Info
No
Hypothethical protein found?
Yes
Save the Hit
Back to BLAST Cascade
NOT successful
identity < 98%
Report of Current DB, Alignment Info
Comparison for Similar Hits Detail (4)
qpercent & qpercent > 80%
Similar HITS Swissprot
Similar HITS SpTrembl
Similar HITS Refseq
60% - 85% identity -> similar
40% - 60% identity -> weakly similar
20% - 40% identity -> putative
Merge Results
Select the BEST HIT

Figure 2. Task flow of Protsweep.

In principle, it is possible to decide the significance of a match upon its *E*-value. However, there are a few complications such as that there is no analytical results available for accurately determining *E*-values for gapped alignments, especially profile HMM alignments. We use as threshold the trusted cut-off value (TC) existing for each PFAM family. This value is the lowest score for sequences included in the family (e.g. in the full alignment). Therefore, we consider a hit very significant if scores better than the trusted cut-off and at the same time has a significant *E*-value. In the case of SCOP, individual protein families are described by several HMMs. We use the SCOP filtering mechanism to look for consistency in the HMMScan output, and filtering out inconsistent hits. In the case of SMART we use only the *E*-value. For each of the protein/domain databases used, we have established different thresholds and rules.

Afterwards DomainSweep takes all true positive hits of all individual database searches for further data interpretation. Domain hits are listed as 'significant':

(i) If two or more hits belong to the same INTERPRO family. The task compares all true positive hits of the different protein family databases grouping together those hits, which are members of the same INTERPRO family/domain.
(ii) If the motif shows the same order as described in PRINTS or BLOCKS. Both databases characterize a protein family with a group of highly conserved motifs/segments in a well-defined order. The task compares the order of the identified true positive hits with the order described in the corresponding PRINTS or BLOCKS entry. Only hits in correct order are accepted.

All other true positive hits are listed as 'putative' (Figure 3).

It is clear that any automatically produced sequence analysis implies a reasonable compromise between sensitivity and selectivity, and that no ideal recognition threshold exists that would allow for perfect separation of true and false similarities. Our thresholds tend to be rather conservative and stringent and thus the possibility of extending false positives is very limited.

The output in the web consists of two groups of graphs, those corresponding to the significant and putative hits, and one table output containing all reported protein domains (Supplementary Figure 2). The graphical outputs display for each 'significant' or 'putative' hit a cartoon of

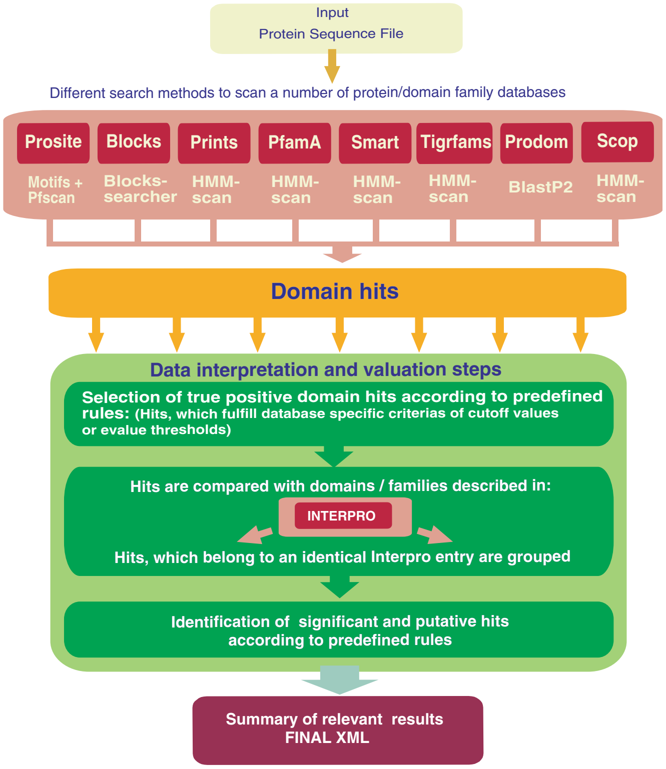

Input
Protein Sequence File
Different search methods to scan a number of protein/domain family databases
Prosite
Blocks
Prints
PfamA
Smart
Tigrfams
Prodom
Scop
Motifs + Pfscan
Blocks-searcher
HMM-scan
HMM-scan
HMM-scan
HMM-scan
BlastP2
HMM-scan
Domain hits
Data interpretation and valuation steps
Selection of true positive domain hits according to predefined rules: (Hits, which fulfill database specific criterias of cutoff values or evalue thresholds)
Hits are compared with domains / families described in:
INTERPRO
Hits, which belong to an identical Interpro entry are grouped
Identification of significant and putative hits according to predefined rules
Summary of relevant results
FINAL XML

Figure 3. DomainSweep overall processing flowchart.

the sequence with the domain corresponding to the match, the hit ID, description, begin, end and Gene Ontology (GO) annotation. The user has immediate access to all complete application outputs and database entries (via hyperlinks) by clicking on the corresponding part of the picture. At the bottom of each graph there is a link to the task explanatory legend. The table output contains all hits, IDs, descriptions and links to the original output. The XML output containing all the generated information is available via hyperlink at the bottom of the task output.

2DSweep identifies the structural domains in the protein and therefore aids in finding structural elements. It reports on predictions for alpha-helix, beta-strand, coiled-coil and helix-turn-helix motifs, transmembrane regions, signal sequences, hydrophobicity, antigenicity, protease cleavage sites and more.

When predicting the secondary structure of a protein, it is useful to exploit the features of several available prediction algorithms rather than to rely on a single program. Unfortunately, combining prediction methods on a large scale is complicated by the fact that prediction programs have very different input requirements and output formats. Some of them perform much better when they have a multiple sequence alignment covering different degrees of similarity as input instead of a single sequence.

We have developed MSFGenerator, a program, which creates a multiple sequence alignment for a single protein sequence according to user, defined rules (Supplementary Data MSF). It performs a BLAST search against a non-redundant protein database following different strategies that will generate different kind of alignments (Supplementary Data MSF, Figure 4). The output of MSFGenerator is an alignment in MSF format (multiple sequence file). The generated MSF will be used as input for four different structure prediction programs: PsiPred (18), Jnet (19), Prof (20), and DSC (21). Each derives its prediction using a different heuristic. PsiPred is a two-stage neural network that bases its prediction on position specific scoring matrices, Jnet is a neural network method that works by utilizing an alignment as input, alongside Psiblast (22) and HMM profiles. Prof is a classifier that combines linear discriminations and neural networks. DSC is based on decomposing secondary structure prediction into basic concepts and then uses simple and linear statistical methods to combine them. Since DSC is known to perform worse than the other prediction methods employed in 2Dsweep, the usage of DSC is optional.

As a second concept, 2DSweep searches for DSSP (Definition of Secondary Structure of the Protein, (23) annotation for the input protein. 2DSweep runs a Blast against the PDB database. For all local alignments found it extracts secondary structure elements (if any) from the structure definition of the DSSP database. If there is more than one element covering the same sequence region, 2DSweep uses a simple majority vote to determine the structure at each position. The result of this procedure is shown together with the prediction of the different secondary structure prediction tools. Additionally, 2DSweep shows several other common measures of secondary structure. First, the distribution of small, charged and hydrophilic amino acids are shown and probable antigenic regions are indicated.

Furthermore, the task searches for transmembrane helices and intervening loop regions using four different methods: TmHmm (24), DAS (25), TMap (26) and TmPred. In eukaryotic protein sequences, it additionally searches for signal peptides. Finally, information is given about molecular weight, isoelectric point, the distribution of protease cleavage-sites, and the possible sub-cellular localization of the protein.

The web output of 2DSweep (Supplementary Figure 3) is divided in five sections: (i) General Information, (ii) Secondary structure, (iii) Features and (iv) Cleavage sites. The information provided in each of these sections is shown in Figure 4 and Supplementary Table 3. The complete results can be viewed by clicking on the corresponding part of the picture. At the bottom of each graph there is a link to the corresponding explanatory legend. As in the other tasks the XML output containing all the generated information is available via a hyperlink at the bottom of the task output.

IMPLEMENTATION

These servers have been implemented using the W3H task framework (27), which allows the execution of compound jobs using work and data flow descriptions in a heterogeneous bioinformatics environment using meta-data information. The system regulates the dataflow by

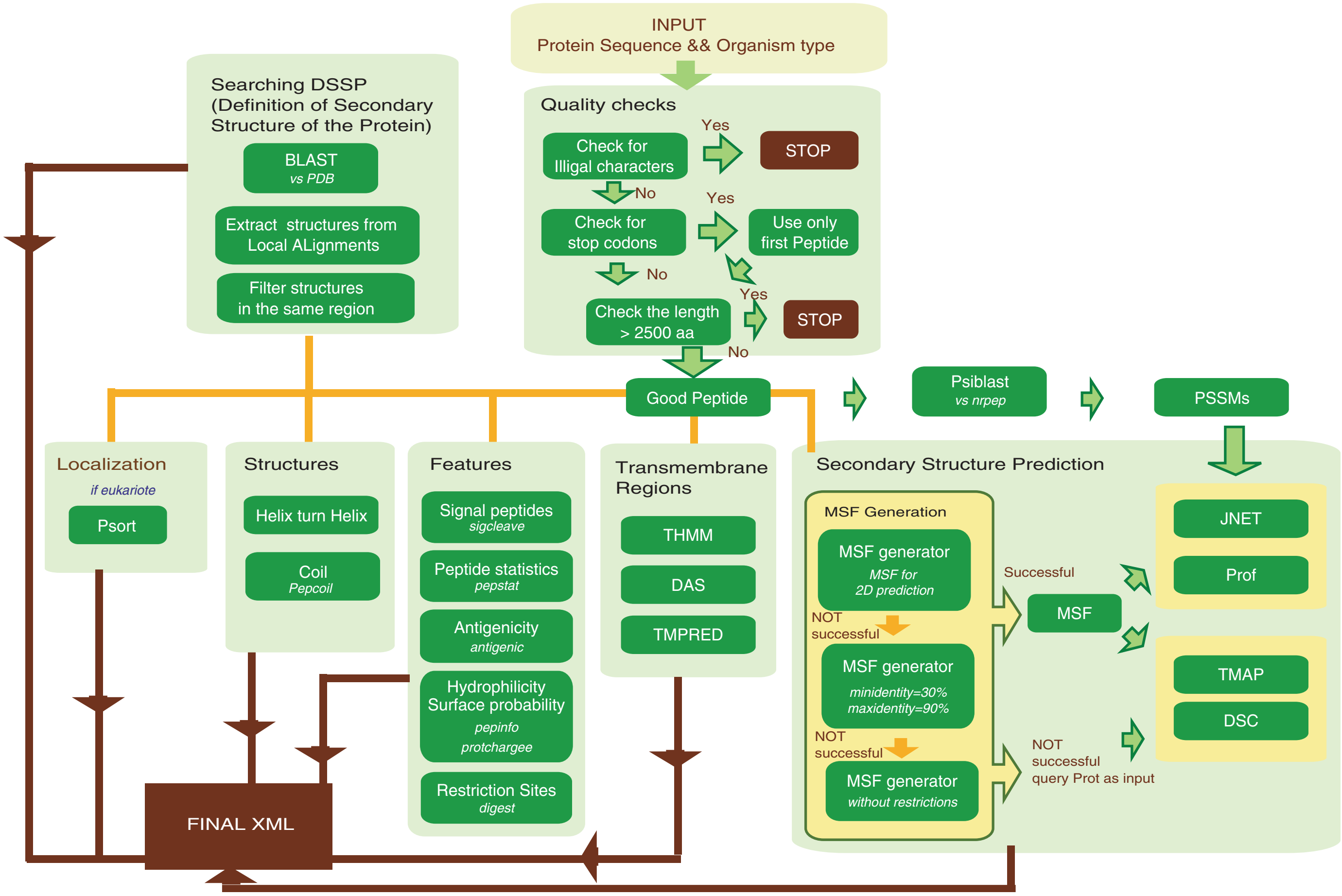


Figure 4. 2DSweep flowchart.

specifying dependency rules between the used applications in the meta-data, which allows the design of high complexity bioinformatics tasks, and stores the results of the different applications together with the new results computed during the process.

The final output of the task is an XML file which contains all relevant information generated. The XML information is transformed by means of W2H's (28) post-processing mechanism into an HTML page for the task report using the Extensible Style-sheet Language Transformations (XSLT; http://www.w3.org/TR/xslt for facilitating a final visual inspection of the results. Furthermore, the XML output can be also required and used for further analysis (i.e. direct integration in user's databases, additional pipeline analysis). All public databases used by these servers are installed under the Sequence Retrieval System (SRS) at the DKFZ (29). The DKFZ SRS server contains more than 500 databases that are automatically updated whenever new releases become available; this means that the webservers will be using the very last version of each database.

The use of this integrated approach provides great flexibility and extensibility of the process. Therefore, as new and improved algorithms and methodologies are developed, they are incorporated into the protein analysis process without having to redesign the entire task. It is also possible to incorporate specific sets of databases as they become available, and to implement arbitrary configuration parameters.

OUTLINE

The development of the three pipelines presented here, has been user-driven from the beginning. Their functionalities are continually being updated and extended in response to requests and suggestions emerging from our core users like LIFEDB (30,31), where these servers are actively used in their protein analysis and annotation.

We are currently developing checks especially through the application of filtering strategies and algorithms that will take into account the relationships between domain structure and homology searches. At the moment we are starting to develop a filtering system for the homology searches results taking into account the different quality of annotation in different protein databases with the idea to assign confidence levels and cross-checking results between tasks. We are additionally working on the implementation of directed text mining using the keywords of the proteins description.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We want to thank Stefan Wiemann, Alex Mehrle and Ingo Schupp for fruitful discussions and suggestions during the developmental phase and for intense testing in the relase phase. Bernhard Heinzel for his help in programming 2DSweep, Ulrich Pfisterer, Katharina Willmann, Zsuzsanna Varga and Jennifer Honek for benchmarks and testing during the entire process. The program of 'Retorno de Investigadores de la Junta de Andalucia' financed Dr C. del Val. Funding to pay the Open Access publication charges for this article was provided by Dep. Molecular Biophysics, Deutsches Krebsforschungszentrum.

Conflict of interest statement. None declared.

REFERENCES

1. Wiemann,S., Weil,B., Wellenreuther,R., Gassenhuber,J., Glassl,S., Ansorge,W., Bocher,M., Blocker,H., Bauersachs,S. *et al.* (2001) Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.*, **11**, 422–435.
2. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones.
3. Kikuno,R., Nagase,T., Waki,M. and Ohara,O. (2002) HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.*, **30**, 166–168.
4. Wu,C.H., Huang,H., Yeh,L.S. and Barker,W.C. (2003) Protein family classification and functional annotation. *Comput. Biol. Chem.*, **27**, 37–47.
5. Larsson,T.P., Murray,C.G., Hill,T., Fredriksson,R. and Schioth,H.B. (2005) Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery. *FEBS Lett.*, **579**, 690–698.
6. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
7. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
8. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
9. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
10. Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.
11. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
12. Henikoff,S., Henikoff,J.G. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.
13. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
14. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
15. Ponting,C.P., Schultz,J., Milpetz,F. and Bork,P. (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.*, **27**, 229–232.
16. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
17. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C., Murzin,A.G., Lo Conte,L., Brenner,S.E., Hubbard,T.J. *et al.* (2004) SCOP database in 2004: refinements integrate structure and sequence family data.
18. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
19. Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
20. Ouali,M. and King,R.D. (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.*, **9**, 1162–1176.
21. King,R.D. and Sternberg,M.J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.*, **5**, 2298–2310.
22. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Kabsch,W. and Sander,C. (1985) Identical pentapeptides with different backbones. *Nature*, **317**, 207.
24. Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
25. Cserzo,M., Eisenhaber,F., Eisenhaber,B. and Simon,I. (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics*, **20**, 136–137.
26. Persson,B. and Argos,P. (1994) Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.*, **237**, 182–192.
27. Ernst,P., Glatting,K.H. and Suhai,S. (2003) A task framework for the web interface W2H. *Bioinformatics*, **19**, 278–282.
28. Senger,M., Flores,T., Glatting,K., Ernst,P., Hotz-Wagenblatt,A. and Suhai,S. (1998) W2H: WWW interface to the GCG sequence analysis package. *Bioinformatics*, **14**, 452–457.
29. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
30. Bannasch,D., Mehrle,A., Glatting,K., Pepperkok,R., Poustka,A. and Wiemann,S. (2004) LIFEdb: a database for functional genomics experiments integrating information from extermal sources, and serving as a sample tracking system. *Nucleic Acid Res.*, **32.**
31. Mehrle,A., Rosenfelder,H., Schupp,I., del Val,C., Arlt,D., Hahne,F., Bechtel,S., Simpson,J., Hofmann,O. *et al.* (2006) The LIFEdb database in 2006. *Nucleic Acids Res.*, **34**, D415–D418.