Chapter 25

The Recipe for Protein Sequence-Based Function Prediction and Its Implementation in the ANNOTATOR Software Environment

Birgit Eisenhaber, Durga Kuchibhatla, Westley Sherman, Fernanda L. Sirota, Igor N. Berezovsky, Wing-Cheong Wong, and Frank Eisenhaber

Abstract

As biomolecular sequencing is becoming the main technique in life sciences, functional interpretation of sequences in terms of biomolecular mechanisms with in silico approaches is getting increasingly significant. Function prediction tools are most powerful for protein-coding sequences; yet, the concepts and technologies used for this purpose are not well reflected in bioinformatics textbooks. Notably, protein sequences typically consist of globular domains and non-globular segments. The two types of regions require cardinally different approaches for function prediction. Whereas the former are classic targets for homology-inspired function transfer based on remnant, yet statistically significant sequence similarity to other, characterized sequences, the latter type of regions are characterized by compositional bias or simple, repetitive patterns and require lexical analysis and/or empirical sequence pattern–function correlations. The recipe for function prediction recommends first to find all types of non-globular segments and, then, to subject the remaining query sequence to sequence similarity searches. We provide an updated description of the ANNOTATOR software environment as an advanced example of a software platform that facilitates protein sequence-based function prediction.

Key words Protein sequence analysis, Protein function prediction, Globular domain, Non-globular segment, Genome annotation, ANNOTATOR

1 Introduction

Advances in sequencing technology have driven costs to such low levels that DNA, genome, and RNA sequencing have become the main research technologies in life sciences and they get applied in various context not necessarily because these methods are the most appropriate ones for the task but they have become the most accurate, affordable methods and they are also increasingly generally available; so, people just do it [1–3]. The results are heaps of sequence data where only a minor fraction is functionally understood and interpreted.

Oliviero Carugo and Frank Eisenhaber (eds.), *Data Mining Techniques for the Life Sciences*, Methods in Molecular Biology, vol. 1415, DOI 10.1007/978-1-4939-3572-7_25, © Springer Science+Business Media New York 2016

The issue is best illustrated by the number of genes that remain without function despite having been sequenced longer than a decade ago. For example, among the almost 7000 genes of the yeast *Saccharomyces cerevisae*, more than 1000 still awaited their functional characterization in 2007 [4] and little has changed since then. To note, the yeast genome has been available since 1997 and yeast is one of the best studied organisms. In human, just 1.5 % of the genome is protein coding with 20000–25000 genes and about half of them lack function description at the molecular and/or cellular level. The remaining genome is known also to be functionally significant; yet, the molecular mechanisms involving the various non-coding transcripts are largely unknown. The classical route to functional characterization involving experimental methods from the genetic and biochemical toolbox like specific knock-outs, targeted mutations, and a battery of biochemical assays is laborious, time-consuming, and expensive. Thus, concepts, approaches, and tools for sequence-based function prediction are very much needed to guide experimental biological and biomedical discovery-oriented research along promising hypotheses.

As proteins are known to be for a large variety of biological functions and mechanisms, hints about their function are especially valuable. Notably, protein function is described within hierarchical concept [5]. The protein's molecular function set are the functional opportunities that a protein provides for interactions with other molecular players, its binding capacities and enzymatic activities, the range of conformational changes and posttranslational modifications. A subset of these molecular functions becomes actually relevant in the biological context at the cellular level, in biomolecular mechanisms such as metabolic pathways, signaling cascades or supramolecular complexes together with other biomacromolecules (cellular function). Finally, a protein's phenotypic function is its result of cooperation with various biomolecular mechanisms under certain environmental conditions.

As experimental characterization of an uncharacterized protein's function is time-consuming, costly and risky and as researchers follow the pressure toward short-term publishable results, experimentalists tend to concentrate on very few widely studied gene examples which apparently show the greatest promise for the development of drugs, while ignoring a treasure trove of uncharacterized ones that might hold the key to completely new pathways. In-silico sequence analysis aimed at structure/function prediction can become extremely helpful in generating trusted functional hypotheses. In principle, it is fast (up to a few months of effort) and, with the exception of some compute-intensive homology search heuristics [6], it has become affordable for even small-scale research operations independently or, the easiest way, in collaboration with an internationally well-known sequence-analytic research group.

This is not to say that in-silico analysis generates a function discovery for any query sequence or assesses the effect of any mutation in a functionally characterized gene. Nothing is farther from the truth. Yet, if properly applied, the set of sequence-analytic methods provides options and insights that are orthogonal to those provided by other, especially experimental methods and, with some luck, they can deliver the critical information for the path to the success [7]. The field of function prediction from protein sequence is still evolving. Only for some fraction of the uncharacterized sequence targets, predictions that provide useful hints can be made; yet, with a growing body of biological sequences and other life science knowledge, the number of such targets increases. For example, more sequences imply a denser sequence space and greater chances of success for homology-based function prediction as the recent breakthrough for Gaa1/Gpaa1, a subunit of the transamidase complex with predicted metallo-peptide-synthetase activity, has demonstrated [8, 9]. As a matter of fact, function prediction from sequence has made bioinformatics center stage in life science and exercises its influence in all research fields. Further examples are provided in these references [10–13].

It should be noted that certain prediction algorithms, especially many among those for predicting functional features in non-globular segments, are plagued by high false-positive rates. Nevertheless, they might be not completely useless. This is especially true if they are applied in conjunction with experimental screening methods with large lists of genes relevant for certain physiological situations as output. Gene expression studies at the RNA or protein level are typical examples. Function prediction tools can serve as filters for dramatically reducing the list, thus, helping to select gene targets for further experimental follow-up studies.

Taken together, the number and the order of structural and functional segments in a protein sequence are called the sequence architecture (historically, it was just the order of globular domains in the sequence). The sequence architecture is computed by using a variety of sequence-analytic tools over the query sequence. One of the practical problems is that, for each query sequence, it is desirable to apply all known good prediction tools (those with good prediction accuracy) with the hope that at least some of them generate useful information for the query. There are about 40 of such tools available at this time point and many of them need to be run with several parameter sets. Historically, bioinformatics researchers provide their individual prediction algorithms as downloadable programs or web-based services. While generally useful for very specific questions, the input and output formats of these programs tend to be incompatible. It is a considerable workload to feed all the programs and web services with suitable input and to collect the output. Further, the total output for a single protein

with ~1000 amino acids can run into GBs and just reading and extracting the useful annotation correctly can become difficult.

These problems multiply with the number of queries to study. Large sequencing projects require the annotation of thousands of proteins. The answer to this challenge is the implementation of script-based annotation pipelines that chain together several prediction tools and perform the necessary reformatting of inputs and outputs with web-accessible visualization of final results. While being adequate for a particular project, these pipelines lack the flexibility of applying modified sets of algorithms with change of task. An alternative are workflow tools that allow for the integration of a large number of individual prediction algorithms while presenting the results through a unified visual interface and keeping them persisted as well as traceable to the original raw output of sequence-analytic programs. The ANNOTATOR [13, 14] and its derivatives ANNIE [15], a fast tool for generating sequence architectures, and HPMV [16], a tool for mapping and evaluating sequence mutations with regard to their effect on sequence architecture, are representatives of this advanced class of sequence analysis frameworks.

2 Concepts in Protein Sequence Analysis and Function Prediction

The most basic concept in protein sequence studies is centered on the idea of segment-based analysis. Proteins are known to consist of structural and functional modules [17], of segments that have structural properties relatively independent from the rest of the protein and that carry an own molecular function. The final interpretation of protein function arises as a synthesis of the individual segment's functions.

Notably, there are two types of segments. Protein sequences typically consist of globular domains and non-globular segments [18–21]. The two types of regions require cardinally different approaches for function prediction. Sequence segments for globular domains have typically a mixed, lexically complex protein sequence with a balanced composition of hydrophobic and hydrophilic residues where the former tend to compose the tightly packed core and the latter form the surface of the globule [17, 21]. Functionally, globular domains with their unique 3D structure offer enzymatic and docking sites. Since the hydrophobic sequence pattern is characteristic for the fold, even a remnant sequence similarity without any sequence identity just with coincidence of the polar/non-polar succession is strongly indicative for fold similarity, common evolutionary origin, and similarity of function. Therefore, function annotation transfer justified by the sequence homology concept is possible within families of such protein segments that have statistically significant sequence similarity [22].

In contrast, non-globular regions have typically a biased amino composition or a simple, repetitive pattern (e.g., $[GXP]_n$ in the case of collagen) due to physical constraints as a result of conformational flexibility in an aqueous environment, membrane embedding, or fibrillar structure [22–24]. As a consequence, sequence similarity is not necessarily a sign of common evolutionary origin and common function. Non-globular regions carry important functions hosting sequence signals for intracellular translocation (targeting peptides) and posttranslational modifications [24], serving as linkers or fitting sites for interactions. For their functional study, lexical analysis is required and the application of certain types of pattern–function correlation schemes is recommended. Thus, non-globular features require many dozens of tools to locate them in the sequence whereas globular domains are functionally annotated uniformly with a battery of sequence similarity search programs.

Correspondingly, the recipe for function prediction recommends first to find all types of non-globular segments with all available tools for that purpose (step one) and, then to subtract these non-globular regions from the query sequence [21]. The remaining sequence is then considered to consist of globular domains. Since most sequence similarity programs have an upper limit in the number of similar protein sequences in the output, it might happen that sequences corresponding to domains very frequent in the sequence databases overwhelm the output and certain section of the sequence are not covered by hits of sequence-similarity searching programs at all, even if they exist in the database. Therefore, it is recommended to check for the occurrence of well-studied domains in the remainder of the query sequence (step two). A variety of protein domain libraries is available for this purpose.

After subtracting the sequence segments that represent known domains from the query, the final remainder is believed to consist of new domains not represented in the domain libraries. At this time point, the actual sequence similarity search tools have to kick in to collect the family of statistically similar sequence segments (step three). The hope is that at least one of the sequences found was previously functionally characterized so that it becomes possible to speculate about the function of this domain as, for example, in [25–29].

The existence of homologous sequences with experimentally determined three-dimensional structures opens the possibility to use them as templates for computationally modeling the 3D structure of the query sequence. Determining the evolutionary conservation of individual residues and, then, projecting these values onto the modeled 3D structure can give valuable hints as to interaction interfaces or catalytic sites. This approach was useful to provide crucial insights into mechanisms for the development of drug resistance as the example of the H1N1-Neuraminidase shows [30] but also in other contexts [31]. 3D structure modeling within

the homology concept is a complex task with many own parameters that is best executed outside of the ANNOTATOR, for example with the MODELLER tool [32–35].

3 ANNOTATOR: The Integration of Protein Sequence-Analytic Tools

The ANNOTATOR software environment is actively being developed at the Bioinformatics Institute, A*STAR (http://www.annotator.org). This software environment implements many of the features discussed above. Biological objects are represented in a unified data model and long-term persistence in a relational database is supplied by an object-relational mapping layer. Data to be analyzed can be provided in different formats ranging from web-based forms, FASTA formatted flat files to remote import over a SOAP interface. This interface provides also an opportunity for other programs to use the ANNOTATOR as a compute engine and process the prediction results in their own unique way (e.g., ANNIE [15] and HPMV [16]).

At the moment, about 40 external sequence-analytic algorithms from own developments or from the academic community are integrated using a plugin-style mechanism and can be applied to uploaded sets of sequences (see the large Table 1 for details). The display of applicable algorithms follows the three-step recipe described above. Integrated algorithms that execute complex tasks such as ortholog or sequence family searches constitute a further group of algorithms. Finally, the ANNOTATOR provides tools to manage sequence sets (alignments and sequence clustering).

1. Searching for non-globular domains.
   (a) Tests for segments with amino acid compositional bias and disordered regions.
   (b) Tests for sequence complexity.
   (c) Prediction of posttranslational modifications.
   (d) Prediction of targeting signals.
   (e) Prediction of membrane-embedded regions.
   (f) Prediction of fibrillar structures and secondary structure.
2. Searching for well-studied globular domains.
   (a) Searches in protein domain libraries.
   (b) Tests for small motifs.
   (c) Searches for repeated sequence segments.
3. Searching for families of sequence segments corresponding to new domains.
4. Integrated algorithms.

Table 1
Algorithms and sequence-analytic tools integrated in the ANNOTATOR

Algorithm	Description	Standard-parameters
Non-globular regions		
Compositional bias		
CAST [48, 49]	The CAST algorithm is based on multiple-pass Smith–Waterman comparison of the query sequence against 20 homopolymers with infinite gap penalties. The detection of low-complexity regions is highly specific for single residue types. CAST might be used as alternative to SEG for masking compositionally biased regions in queries prior to database-wide sequence comparisons such as BLAST	Threshold = 40
DisEMBL [50, 51]	DisEMBL is a computational tool for prediction of disordered/unstructured regions within a protein sequence. The method is based on artificial neural networks trained for predicting three different definitions of disorder: loops/coils, hot-loops, and Remark-465 (missing coordinates)	Minimum peak width = 8 Maximum join distance = 4, coils threshold = 1.2 Remark465 threshold = 1.2 Hot loops threshold = 1.4
GlobPlot 1.2 [52]	The GlobPlot algorithm measures and displays the propensity of protein sequences to be ordered or disordered. It is a simple approach based on a running sum of the propensity for amino acids to be in an ordered or disordered state	Minimum peak width (disorder prediction) = 8 Minimum peak width (globular domain hunting) = 8 Maximum join distance (disorder prediction) = 4 Maximum join distance (globular domain hunting) = 4, Smoothing frame = 8 (Savitzky–Golay) Propensity set = Russell/Linding
IUPred [53, 54]	IUPred is a prediction method for recognizing ordered and intrinsically unstructured/disordered regions in proteins. It is based on estimating the capacity of polypeptides to form stabilizing contacts. The underlying assumption is that globular proteins make a large number of inter-residue interactions, whereas intrinsically unstructured/disordered regions have special amino acid compositions not allowing sufficient favorable interactions to form a stable tertiary structure	Long disorder sequential neighborhood = 100aa Short disorder sequential neighborhood = 25aa Structured regions minimum size = 30aa

(continued)

Table 1
(continued)

Algorithm	Description	Standard-parameters
SAPS [55]	SAPS evaluates a wide variety of protein sequence properties by statistical criteria. Properties include global compositional biases, local clustering of different residue types (e.g., charged residues, hydrophobic residues, Ser/Thr), long runs of charged or uncharged residues, periodic patterns, counts and distribution of homooligopeptides, and unusual spacings between particular residue types	The residue composition of the input protein sequence is evaluated relative to SWISS-PROT (from the year of SAPS publication 1992) by default
XNU [56, 57]	XNU identifies self-redundancy within a protein sequence classified into two categories: internal repeats and intrinsic repeats. Internal repeats are the tandem arrangements of discrete units (which can also be globular domains like IG, EGF, and other typical repeat domains). Intrinsic repeats are the compositionally biased segments of a small number of distinct amino acids with no clear repeating pattern. These repeats are identified on a dot-plot matrix of self-comparison of the query sequence by scoring the local similarity with a PAM matrix and estimating the statistical significance of the score	Probability cutoff = 0.01 Search-width = 10 Scoring matrix = PAM120
DisoPred [58]	DISOPRED predicts protein disorder DISOPRED2 was trained on a set of sequences with high-resolution X-ray structures where residues appear in the sequence records but not in the coordinates (missing electron density). Sequence profile was generated using PSI-BLAST and the data were used to train linear supportvector machines	False-positive threshold = 5 % Min length of detected region = 2 Max gap within region = 2 Subject sets: NCBI non-redundant protein set PDB PDB and UniRef90 UniRef90 sequence clusters
Sequence complexity		
SEG [59–62]	Low complexity regions (LCRs) represent sequences of very non-random composition ("simple sequences", "compositionally biased regions"). They are abundant in natural sequences. SEG is a program providing a measure of compositional complexity of a segment of sequence and divides sequences into contrasting segments of low-complexity and high-complexity. Typically, globular domains have higher sequence complexity than fibrillar or conformationally disordered protein segments	Annotator provides three parameter sets: (1) SEG12: Window Size = 12; Locut = 2.2; Hicut = 2.5 (2) SEG25: Window Size = 25; Locut = 3.0; Hicut = 3.3 (3) SEG45: Window Size = 45; Locut = 3.4; Hicut = 3.75

Posttranslational modifications		
Big PI [63–66]	Posttranslational modification with a glycosylphosphatidylinositol (GPI) lipid anchor is an important mechanism for tethering proteins of eukaryotic organisms and their viruses to cellular membranes Big-Pi is a program for the prediction of suitable candidates for GPI lipid anchoring. It identifies the cleavage site in the C-terminally located GPI signal. The predictive accuracy is estimated to be clearly over 80 % for metazoan, plant, and fungal proteins and almost 80 % for protozoan proteins. The false-positive prediction rate is estimated to be in the range of 0.1 %	Learning sets: Big PI: Metazoa Protozoa Big Pi3.2: Metazoa Protozoa Fungi Viridiplantae
MyrPS/NMT [18, 67–70]	Myristoylation is a lipid modification at the N-terminus of eukaryotic and viral proteins. The enzyme myristoylCoA:protein N-myristoyltransferase (NMT) recognizes certain characteristics within the N-termini of substrate proteins and finally attaches the lipid moiety to a required N-terminal glycine By analysis of known substrate protein sequences and kinetic data, the motif for N-terminal (glycine) myristoylation was refined and three motif regions were identified: region 1 (positions 1–6) fitting the binding pocket, region 2 (positions 7–10) interacting with the NMT's surface at the mouth of the catalytic cavity, and region 3 (positions 11–17) comprising a hydrophilic linker. Each region was characterized by specific requirements concerning volume compensations, polarity, flexibility parameters, and other typical properties of amino acid side chains. Additionally, evolutionary shifts between lower and higher eukaryotic NMT sequences resulting in taxon-specific substrate preferences were observed. This motif description was implemented in a function that scores query sequences for suitability as NMT substrates and the scores are also translated into probabilities of false-positive predictions. .	Parameter set: Non-fungal eukaryotes and their viruses
PrePS/Prenylation-FT [71–73]	Prenylation refers to the posttranslational modification of proteins with isoprenyl anchors. This predictor aims to model the substrate–enzyme interaction based on refinement of the recognition motif of the eukaryotic enzyme farnesyltransferase (FT) Motif information has been extracted from sets of known substrates (learning sets). Specific scoring functions have been created utilizing both sequence as well as physical property profiles including interpositional correlations and accounting for partially overlapping substrate specificities with other prenyltransferases	None
PrePS/Prenylation-GGT1 [71–73]	This is a prenylation predictor similar to Prenylation-FT and Prenylation-GGT2. It aims to model the substrate–enzyme interaction based on refinement of the recognition motif of the eukaryotic enzyme geranylgeranyltransferase 1 (GGT1)	None

(continued)

Table 1
(continued)

Algorithm	Description	Standard-parameters
PrePS/Prenylation-GGT2 [71–73]	This is a prenylation predictor similar to Prenylation-FT and Prenylation-GGT1. It aims to model the substrate–enzyme interaction based on refinement of the recognition motif of the eukaryotic enzyme geranylgeranyltransferase 2 (GGT2 or RabGGT)	None
Targeting signals		
PeroxyPS/PTS1 [74, 75]	Peroxisomal matrix proteins have to be imported into their target organelle posttranslationally. The major translocation pathway depends on a C-terminal targeting signal, termed PTS1. The PTS1 signal predictor finds proteins with a C-terminus appropriate for peroxisomal import. It is capable of recognizing potential PTS1s in query sequences	Prediction function = General
SIGCLEAVE [76, 77]	Signal peptide-mediated translocation of nascent proteins from the cytoplasm across the endoplasmic reticulum membrane is a major export mechanism in eukaryotes. In prokaryotes, signal peptides mediate translocation across the cellular membrane. SigCleave is a program (originally part of the EGCG molecular biology package) to predict signal sequences. It identifies the cleavage site between a signal sequence and the mature exported protein based on the von Heijne (1986) algorithm. The predictive accuracy is estimated to be 75–80 % for both prokaryotic and eukaryotic proteins (Menne KM, Hermjakob H, Apweiler R (2000) Bioinformaticss 16,741–2)	Taxon: prokaryotes and eukaryotes Threshold: 3.5
SignalP-3.0 [78–80]	SIGNALP predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive bacteria, gram-negative bacteria, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks. It also incorporates predictions done by a hidden Markov model specifically designed to distinguish between signal peptides, non-secretory proteins, and signal anchors (signal peptides that are not cleaved, for eukaryotes only)	Taxon: all available taxa
Membrane-embedded regions		
DAS-TMfilter [81, 82]	The method discriminates between genuine TM and non-TM queries than the location of the TM regions is predicted when it is appropriate. The tool is based on the “Dense Alignment Surface” algorithm. The estimated efficiency of the method is around 95 % in terms of the location of the TM segments and 99 % in terms of the type of the query	Quality Cutoff: 0.72

HMMTOP 2.0 [83]	The tool implements a Hidden-Markov Model to predict TM protein topology. The engine uses a five-state model: TM helix (H), inner and outer helix tails (i,o), inner and outer loops (I,O). The predictive power of the method is around 95 %	No major adjustable parameters
PHOBIUS [84, 85]	The predictor is based on a hidden Markov model (HMM) that models the different sequence regions of a signal peptide and the different regions of a transmembrane protein in a series of interconnected states	No major adjustable parameters
TMHMM [86, 87]	TMHMM is a membrane protein topology prediction method based on a hidden Markov model. It can discriminate between soluble and membrane proteins with both specificity and sensitivity better than 99 %, although the accuracy drops when signal peptides are present	No major adjustable parameters
TOPPRED [88, 89]	TOPPRED predicts the location of the TM segments in the query using one of the three popular hydrophobicity scale. The topology of the sequence also predicted based on the "positive inside" rule. The predictive power of the method is moderate	Peak Cutoff = 1.0 Organism: Metazoa Protozoa
TM-complexity [22, 23, 47]	Each predicted transmembrane segment is classified with regard to sequence complexity as simple (simple hydrophobic anchor), twilight, or complex (transmembrane region with additional functional or structural role) as defined in [22, 23, 47]	No adjustable parameters
Secondary structure		
ImpCOIL [89–91], *implementation of a slightly modified algorithm by Frank Eisenhaber (2000)*	Coiled coil regions in proteins are bent *alpha*-helices that are packed together in dimer, trimer, or tetramer arrangements. The small docking angle of the helix packing (almost parallel or antiparallel packing) is achieved with high helix radii; i.e., leucine residues or other amino acid types with long hydrophobic side chains are placed at the first and fourth ("a" and "d") positions of an heptade repeat. Sequence profiles of typical heptade repeats have been derived by Lupas et al. which are used in this implementation. High scoring segments are predicted to have helical structure involved in coiled coil packings	None
Predator [89, 92, 93]	PREDATOR program combines propensities of long-range interactions (hydrogen bondings) with a nearest neighbor and a statistical approach. (Frishman D, Argos P, (1997) Proteins 27(3), 329–335). The accuracy of a secondary structure prediction is measured by the Q_3 value, which is defined as the overall percentage of the predicted to the observed secondary structures of specific protein sets. The Q_3 value lies between 68 and 71 %. (Cuff JA, Barton GJ (1999) Proteins 34 (4), 508–519)	None

(continued)

Table 1
(continued)

Algorithm	Description	Standard-parameters
SSCP [94, 95]	Secondary structural content is the relative distribution of residues among *alpha*-helix, *beta*-strand, and coil state The SSCP tool predicts the secondary structural content of a query protein from its amino acid composition with two independent regression methods, (a) by ignoring correlations between pairs of amino acid types and (b) by taking them into account. The predicted secondary structural content can be considered only indicative for the query protein since the exact sequence cannot be ignored in secondary structural content prediction	None
Known sequence domains		
HMMER2 [36, 43–45, 96–98]	HMMER2 is based on hmmpfam. It searches libraries of HMMs for known domains and motifs in a query sequence. Available HMM libraries in ANNOTATOR are Pfam, Smart Fragments, Repeats (Miguel Andrade), Smart, Huf-zinc, Pfam-zinc as well as an internal library	*E*-value-cutoff = 0.01 HMMER Database = Pfam
HMMER3 [36, 37, 99–102] Documentation is available at http://hmmer.org/	HMMER3 is based on hmmscan. It searches a library of HMM profiles for known domains in query sequence(s). Available HMMER3 libraries in ANNOTATOR are Pfam, dbCAN, and antiSMASH	*E*-Value-cutoff = 0.01 HMMER3 database = Pfam
ModEnzA [103]	The ModEnzA algorithm is an implementation of the method HMM-ModE for the identification of metabolic enzymes. HMM-ModE is a protocol to generate family-specific HMMs. It first constructs a profile HMM from an alignment of the family's sequences and then uses this model to identify sequences belonging to other classes that score above the default threshold (false positives). Tenfold cross-validation is used to optimize the discrimination threshold score for the model resulting in HMM-T profiles. The advent of fast multiple alignment methods enables the use of the profile alignments to align the true and false-positive sequences, and the resulting alignments are used to modify the emission probabilities in the original model generating HMM-ModE profiles. The same algorithm is implemented for the pre-classified sequence set of ENZYME nomenclature database and named as ModEnzA. The resulting HMM-T and HMM-ModE profiles are used to annotate a novel sequence for possible similarity with the available profiles	Database: ModEnzA_hmmT_Enzyme

IMPALA [39, 46, 104]	IMPALA (Integrating Matrix Profiles And Local Alignments) package (Schaffer et al., 1999) provides tools to compare a query sequence against a library of position-specific scoring matrices (PSSMs) produced by PSI-BLAST (Altschul et al. 1997). It performs a Smith–Waterman calculation between the query and each profile. Using a Smith–Waterman calculation guarantees to find optimal local alignments, but is time-consuming. Being complementary to and sharing algorithmic solutions to statistical problems with PSI-BLAST, IMPALA turns out to be comparable to PSI-BLAST concerning sensitivity and error rate. The databases of PSSMs are courtesy of Yuri I. Wolf and L. Aravind	*E*-value-cutoff = 10 Filter = false Subject set = PSSM aravind105, PSSM wolf1187
HHPRED [41, 42]	HHpred is based on HHsearch—which searches a query HMM (Hidden Markov Model) against databases of HMMs. The original HHpred takes a query sequence (or MSA) and builds up a query HMM using PSI-BLAST which it then passes to HHsearch. Later versions of HHpred use HHblits to build up the query HMM	*E*-value = 0.001 HMM databases: Pfam PDB70 SCOP
HHblits [40]	HHblits (HMM-HMM-based lightning-fast iterative sequence search), a HMM-HMM-based alignment method, builds high quality alignments by converting query sequence(s) or MSAs into query HMM and searching iteratively through uniprot20 or nr20 databases	Subject set: nr20 Match states: 50 Number of iterations : 2 *E*-value: 0.001 Alignment: local
HHsearch [41, 42]	HHsearch, a HMM-HMM comparison method, is used for detecting distant homologous relationships between proteins. HHsearch converts query sequence or MSA into profile HMM and searches through a database of HMMs for homologous proteins. It is often used in homology modeling In ANNOTATOR, HHsearch can search against PDB70 or Pfam databases	Subject set: pdb70 Match states: 50 *E*-value: 0.001 Alignment: local
PROSITE-Profile [105, 106]	The identification of functional or structural domains and protein families with extreme sequence divergence cannot be easily achieved by using patterns detection(See PROSITE). To overcome this limitation, techniques based on weight matrices (also known as profiles) were developed and implemented to detect such proteins or domains	None

(continued)

Table 1
(continued)

Algorithm	Description	Standard-parameters
RPS-Blast [38]	RPS-BLAST (Reverse PSI-BLAST) allows the comparison of a query sequence against a library of position-specific scoring matrices (PSSMs), and can thereby facilitate the classification and functional annotation of a sequence. RPS-BLAST uses an implementation of a BLAST-like algorithm In ANNOTATOR, RPS_Blast can be run against versions of the common domain databases SMART and PFAM (CDD SMART, and CDD Pfam), the orthologous database COG (CDD COG),eukaryotic orthologous database KOG (CDD Kog), CDD Tigr(TIGRFAMs), NCBI Protein Clusters PRK(CDD Prk), PRK subsets CDD Chl(Chloroplast and organelle proteins), CDD Mth(Mitochondrial proteins), CDD Pha(Phage proteins), CDD Pln(Plant-specific, non-chloroplast proteins), CDD Ptz(Protozoan proteins),as well as a compilation of all these (CDD All)	*E*-Value-cutoff = 10 Filter = false Subject set = CDD All
Small sequence motif libraries		
BioMotif-IMPlibrary Documentation on bioMotif can be obtained at: http://www.lpta.univ-montp2.fr/users/menes/bioMotif_pub/bioMotif.article.1c.b.ps	BioMotif is an external program, written by Gerard Mennessier, which can be called from the Annotator. Its aim is to help the user to find motifs within sets of sequences It can be defined as a language, which allows to store as variables, positions, subsequences, along the search path, for further reference. It also includes a large class of functions and several boolean operators	None
ELM-patterns [107]	Short linear peptide motifs are used for cell compartment targeting, protein–protein interaction, regulation by phosphorylation, acetylation, glycosylation, and a host of other posttranslational modifications. ELM is a resource for predicting functional sites in eukaryotic proteins. Putative functional sites are identified by patterns (regular expressions). Sequence comparisons with short motifs are difficult to evaluate because the usual significance assessments are inappropriate. To improve the predictive power, context-based rules and logical filters are applied to reduce the amount of false positives	ELM pattern: All ELM patterns

PROSITE-patterns [105, 106]	Specific amino acid residues which are important for the biological function (catalytic site, binding sites of prosthetic groups, metal ions or molecules, etc.) of a protein group are more conserved than the overall sequence. Patterns (or regular expressions) are a qualitative description of the consensus sequences for these biological relevant residues. In contrast to profiles (PROSITE-Profiles), there is no statistical evaluation. The pattern either matches or does not. PROSITE is an annotated collection of protein motifs which can be searched for matching patterns with the application PPSearch	None
EF-Patterns [108–110]	EF patterns can be used in the function annotation/prediction, where protein function is being inferred as a combination of Elementary Functions described by the patterns. Structural representatives of EF patterns are Elementary Functional Loops—closed loops (or returns of the protein backbone) determined by the polymer nature of the polypeptide chains with a functional signature. The latter is encoded in the position-specific matrix (PSSM) of the EF pattern, describing the relative importance of every position and frequencies of amino acids performing the function, maintaining the EFL, and its interactions with the rest of the protein	*E*-value: 1.0
Repeated sequence domains		
PROSPERO [111]	PROSPERO can compare a sequence to itself, another sequence or a profile, and print all local alignments with p-values less than some user-defined threshold. Thus prospero is ideal for the analysis of repeats within a sequence. Implementation follows advice of Chris Ponting	*E* thresh = 0.1, Matrix = BLOSUM62
DB search		
NCBI-Blast [62, 112, 113]	BLAST (Basic Local Alignment Search Tool) sequence comparison is used for the task of comparing novel proteins with previously characterized ones, or for delineating regions of sequence conservation. Search speed has been increased compared to initial sequence comparison methods by breaking the query and database sequences into fragments called words. Words found to be similar are extended in both directions attempting to construct an alignment with a score higher than a given threshold. Consequently, BLAST reports local alignment as opposed to global alignment	*E*-value-cutoff = 1E-03 Filter = no filtering Subject set = NCBI non-redundant protein set Matrix = BLOSUM62
OMA-Blast [114, 115]	OMA-Blast is used to find the orthologs of the query protein. BLAST is run against OMA-Set to find orthologous groups of proteins	E-value-cutoff = 1E-03 Filter = no filtering Subject set = OMA-Set Matrix = BLOSUM62

(continued)

Table 1
(continued)

Algorithm	Description	Standard-parameters
PSI-Blast [46, 59, 115]	Position-specific iterative BLAST (PSI-BLAST) is a program of the BLAST package that can be used to search sequence databases for distant, but biologically significant relatives of a query sequence. PSI-BLAST starts with a single input protein sequence and initially conducts a simple BLAST search. In a second step, a reduced multiple sequence alignment is constructed from the initial BLAST, with the length corresponding to the query sequence length (gaps inserted into the query sequence are ignored). For each column of the reduced multiple sequence alignment, the observed residue frequencies are determined, and used to generate a profile of weights (Position-Specific Scoring Matrix). This score matrix is used in the next BLAST run (first iteration) The results of the first iteration BLAST are used to modify the profile which can than be applied to further iterations. Optimally, the iterations are expected to converge on a set of sequences	E-value-cutoff = 10 Inclusion-cutoff = 0.001 Filter = false Subject set = NCBI non-redundant protein set Matrix = BLOSUM62 Number of rounds = 10
CSI-Blast [116]	CS BLAST method derives sequence context-specific amino acid similarities from windows of length 13 centered on each residue. A sequence profile for the query sequence is generated using context-specific pseudocounts and then PSI-BLAST is started with this profile CS BLAST is a simple extension of BLAST. PSI-BLAST is extended to the context-specific protein sequence searching, CSI-BLAST, in a similar fashion	*E*-value-cutoff = 10 Inclusion-cutoff = 0.001 Filter = false Subject set = NCBI non-redundant protein set Number of rounds = 10
GLSearch [117, 118]	GLSearch is part of the Fasta36 program suite. It searches a query sequence against a sequence database using an optimal algorithm that requires the entire query to match (global) at least part (local) of the database sequences. For small sequence databases statistics can be calculated using sequence shuffling	*E*-value = 0.001, *E*-value cutoff = 0.001, Min *E*-value = 0.0, Filter = pseg, Matrix = BLOSUM50, Gap-Open = -12, Gap-Extend = -2 Subject Sets: brix-and-nr999 NCBI NR PDB SeqRes UniRef90 UniRef90 Clusters

Integrated		
Prim-Seq-An [14]	Prim-Seq-An ("Primary Sequence Analysis") runs a standard set of algorithms on a sequence of interest	Algorithms: SAPS GlobPlot (Disorder peak = 8 Globular peak = 8 Disorder join = 4 Globular join = 4 Hunting = DIS) CAST (Threshold = 40) SEG (Window Size = 12 Hicut = 2.5 Locut = 2.2) SEG (Window Size = 25 Hicut = 3.3 Locut = 3.0) SEG (Window Size = 45 Hicut = 3.75 Locut = 3.4) big-PI (Learning Set: protozoa) big-PI (Learning Set: metazoa) big-PI3.2 (Learning Set: protozoa) big-PI3.2 (Learning Set: metazoa) big-PI3.2 (Learning Set: fungi) big-PI3.2 (Learning Set: viridiplantae) MyrPS/NMT (Parameter Set: default) MyrPS/NMT (Parameter Set: fungi) PeroxyPS/PTS1 (Function: general) PeroxyPS/PTS1 (Function: metazoan) PeroxyPS/PTS1 (Function: fungi) PrePS/Prenylation-FT PrePS/Prenylation-GGT1 PrePS/Prenylation-GGT2 SIGCLEAVE (

(continued)

Table 1
(continued)

Algorithm	Description	Standard-parameters
		Threshold =3.5 Cell Type = Both) SignalP DAS-TMfilter (Quality Cutoff: 0.72) TMHMM HMMTOP PHOBIUS impCOIL HMMER (E-Value-Cutoff = 0.01 Display-Cutoff = 20.0 against smart_patterns) RPS-BLAST (E-Value-Cutoff = 0.001 Display-Cutoff = 1.0 Filter: false against CDD All) IMPALA (E-Value-Cutoff = 1.0E-5 Display-Cutoff = 5.0 Filter: false against PSSM wolf1187) IMPALA (E-Value-Cutoff: 1.0E-5 Display-Cutoff: 5.0 Filter: false against PSSM aravind105) PROSITE-Profile

Orphan-Search [14]	Orphan-Search determines whether a sequence is an orphan within a specific sequence database	Parameters: SEG-1 (Window Size = 12 Hicut = 2.5 Locut = 2.2) SEG-2 (Window Size = 25 Hicut = 3.3 Locut = 3.0) Coil (Minimum-Length = 25 Marking Type = Mark with Xs Orphan (E-Value-Cutoff = 1e-5 Display-Cutoff = 1e-5 Filter = no Subject Set: brix-and-nr999 NCBI Non-Redundant Protein Set PDB UniRef90 Sequence Clusters Set PDB and UniRef90 Matrix = BLOSUM62
Family-Searcher [6, 14]	Family Searcher is an efficient tool for tracing distant evolutionary relationships involving large protein families. It is an unsupervised, sensitive sequence segment collection heuristic suitable for assembling very large protein families. It is based on fan-like expanding, iterative database searches. Additional criteria like minimal alignment length, overlap with starting sequence segments, finding starting sequences in reciprocal searches, automated filtering for compositional bias, and repetitive patterns are introduced to prevent inclusion of unrelated hits	Blast Type = PSI-BLAST Blast Flags: Blast DB = NCBI NR Inclusion Cutoff = 0.001 *E*-Value Cutoff = 0.1 Rounds = 5 Filter = No Filtering Matrix = Blosum62 Family-Searcher Flags: Substitution E-Value = 1E-8 Grand-Parent Check = true Grand-Parent Check E-Value = 1E-2 Ancestors-Check = false Next Query E-Value Cutoff = 1E-3

(continued)

Table 1
(continued)

Algorithm	Description	Standard-parameters
		Concatenate Hits = true Merge Hits with X = true Clean with SEG = true Clean with Coil = true Window Size = 12 Locut = 2.2 Hicut = 2.5 Max Rounds = 5 Max Blasts per Round = 100 Clustering = No
Orthologue Search [6, 14]	Orthologue Search is an efficient algorithm to identify the orthologs of a protein. This algorithm applies the Reciprocal-Best-Blast-hit approach. It operates on a single seed sequence for each ortholog group and identifies orthologs and inparalogs. It requires a non-redundant multi-species database of proteomes	E-Value Cutoff = 1E-03 Filter = No Filtering Subject Set = NCBI NR Matrix = BLOSUM62 Maximum Number of Alignments = 250
Disan [119]	Disan ("Disorder Analysis") runs a set of disorder predictors with settings that allow consensus and complimentary predictions (e.g., the different predictors have the same false-positive rate)	Algorithms: DISOPRED2 IUPred: long, short CAST DisEmBL: ComilsThreshold, Rem465Threshold, HotloopsThreshold SEG45, SEG25, SEG12 Disorder Analysis Type: Default 5 % FPR—Short and Long Disordered Regions 5 % FPR—Short Disordered Regions 5 % FPR—Long Disordered Regions

		Highest MCC—Short and Long Disordered Regions Highest MCC—Short Disordered Regions Subject Set: UniRef90 Sequence Clusters Set PDB and UniRef90 PDB NCBI Non-Redundant Protein Set brix-and-nr999
Clustering		
MCL Clustering [120, 121]	MCL clustering uses the "Markov Cluster Algorithm". The MCL algorithm is based on the idea that random walks on a graph will infrequently go from one natural cluster to another. By iterating alternating "expansion" and "inflation" operations, the graph is separated into segments (clusters) where there are no longer any paths between segments. MCL clustering takes a set of sequences runs all-against-all BLAST (blastall) and applies the MCL algorithm to the results	Inflation parameter = 5 Clustering scheme = 7
CD-HIT Clustering [121–124]	CD-HIT is a widely used sequence clustering program that is very fast and can handle large sequence databases. It estimates percent identity by counting the number of identical "words" in a pair of sequences. The shared word count for a sequence being clustered is calculated from a look-up table that maps each possible word to the cluster representatives that contain that word	Cluster identity threshold = 0.9 Word size = 5 Length of throw-away-sequences = 10 Tolerance for redundance = 2
Multiple sequence alignment		
T-coffee [125]	T-coffee is broadly based on the progressive approach to multiple alignment. It creates a library of all pairwise sequence alignments. Intermediate alignments are based on the sequences to be aligned next and also on how all of the sequences align with each other	No major adjustable parameters for algorithm itself but a better alignment may result from discarding input sequences that differ significantly from the median input sequence length
Muscle [126, 127]	Muscle applies iterative improvements to the progressive alignments with fast distance estimation using kmer counting, a log-expectation score, and refinement using tree-dependent restricted partitioning	No major adjustable parameters for algorithm itself but a better alignment may result from discarding input sequences that differ significantly from the median input sequence length

(continued)

Table 1
(continued)

Algorithm	Description	Standard-parameters
Probcons [128]	Probcons uses an approach somewhat similar to T-coffee but with the quality of the pairwise alignments calculated using an HMM formalism ("probabilistic consistency"). It also provides iterative post-processing by partitioning the alignment and re-aligning	Consistency reps = 2 Iterative refinement reps = 100 Pre-training reps = 0 And a better alignment may result from discard input sequences that differ significantly from the median input sequence length
Mafft [129–132]	Mafft is based on using Fast Fourier Transforms (FFT) with residue volume and polarity to quickly find homologous regions. It offers a variety of different methods: the original very fast "FFT" methods, "NW" methods that use the Needleman–Wunch algorithm instead of FFT, and newer "INS" methods with iterative refinement based on pairwise alignments	ACCURACY ORIENTED METHODS: L-INS-i (local pairwise alignments) G-INS-i (global pairwise alignments) E-INS-i (for large unalignable regions) SPEED ORIENTED METHODS: FFT-NS-2 (fast; progressive method) FFT-NS-1 (very fast; progressive with a rough guide tree) FFT-NS-2 (2 iterative refinements) FFT-NS-i (i iterative refinements) NW-NS-2 (2 refinements without FFT) NW-NS-i (I refinements without FFT) NW-NS-PartTree-1 (PartTree algorithm)
Others		
Hmmer2 Profile [36]	Hmmer2 Profile constructs a hmm profile for query MSA. It is based on hmmbuild and hmmcalibrate of hmmer2	None
HMMERsearch [36]	HMMERsearch is based on hmmer2 hmmsearch. It takes a query HMM and searches against a sequence database to find similar sequences In ANNOTATOR, hmmsearch can search against NCBI Non-Redundant Protein database, PDB or UniRef90 databases	*E*-value-cutoff: 0.01 Subject Set: NCBI Non-Redundant Protein database

This table provides an overview about all the elementary and integrated prediction and annotation tools available in the ANNOTATOR system. The table is an update from the respective compilation in [13]

5. Sequence sets: Clustering algorithms.
6. Sequence sets: Multiple alignment algorithms.
7. Sequence sets: Miscellaneous algorithms.

Integrated algorithms offer either complex operations over individual sequences or also over sequence sets. The ANNOTATOR provides an integrated algorithm ("Prim-Seq-An") that executes automatically the first two steps of the protein sequence analysis recipe. It tests the query sequence for the occurrence of any non-globular feature as well as for hits by any globular domain or motive database. For this purpose, the complete query sequence is subjected to the full set of respective prediction tools. The results can be viewed in an aggregated interactive cartoon.

The matching of domain models with query sequence segments is, similar to many other sequence-analytic problems, a continuing area of research and, consequently, the ANNOTATOR is subject to continuous change in adopted external algorithms. Domain model matching is mostly performed with HMMER-style [36, 37], other profile-based [38, 39], or profile–profile searches [40–42]. There are issues with the *P*-value statistics applied that have significance for hit selection and that can be improved compared with the original implementation [43]. The sensitivity for remote similarities increases in searches where domain models are reduced to the fold-critical contributions; profile sections corresponding to non-globular parts are advised to be suppressed as in the dissectHMMER concept [44, 45].

Within the third step of the segment-based analysis approach, the identification of distantly related homologs to query sequence segments that remain without match in the preceding two analysis steps is the key task. While tools like PSI-BLAST [46] exist that provide a standard form of iterative family collection, it is often necessary to implement a more sophisticated heuristic to detect weaker links throughout the sequence space. The implementation of such a heuristic might require, among other tasks, the combination of numerous external algorithms such as PSI-BLAST or other similarity search tools with masking of low complexity segments, coiled coils, simple transmembrane regions [23] and other types of non-globular regions, the manipulation of alignments as well as the persistence of intermediate results (e.g., spawning of new similarity searches with sequence hits from previous steps).

Obviously, the mechanism of wrapping an external algorithm would not be sufficient in this case. While the logics of the heuristic could be implemented externally, it would still need access to internal data objects, as well as the ability to submit jobs to a compute-cluster. For this reason, an extension mechanism for the ANNOTATOR was devised which allows for the integration of algorithms that need access to internal mechanisms and data. A typical example for using this extension mechanism to implement a sophisticated search heuristic is the "Family-Searcher", an integrated algorithm that is used to

uncover homology relationships within large superfamilies of protein sequences. Applying this algorithm, the evolutionary relationship between classical mammalian lipases and the human adipose triglyceride lipase (ATGL) was established [6]. For such large sequence families, the amount of data produced when starting with one particular sequence as a seed can easily cross the Terabyte barrier. At the same time, the iterative procedure will spawn the execution of tens of thousands of individual homology searches. It is clearly necessary to have access to a cluster of compute nodes for the heuristic and to have sophisticated software tools for the analysis of the vast output to terminate the task in a reasonable timeframe.

3.1 Visualization

The visualization of results is an important aspect of a sequence analysis system because it allows an expert to gain an immediate condensed overview of possible functional assignments. The ANNOTATOR offers specific visualizers both at the individual sequence as well as at the set level.

The visualizer for an individual sequence projects all regions that have been found to be functionally relevant onto the original sequence. The regions are grouped into panes and are color-coded, which makes it easy to spot consensus among a number of predictors for the same kind of feature (e.g., transmembrane regions that are simple (blue), twilight (yellow-orange), and complex (red) are differently color-coded [23, 47]). Zooming capabilities as well as rulers facilitate the exact localization of relevant amino acids.

The ability to analyze potentially large sets of sequences marks a qualitative step up from the focus on individual proteins. Alternative views of sets of proteins make it possible to find features that are conspicuously more frequent pointing to some interesting property of the sequence set in question. The *histogram view* in the ANNOTATOR is an example of such a view. It displays a diagram where individual features (e.g., domains) are ordered by their abundance within a set of sequences.

Another example is the *taxonomy view*. It shows the taxonomic distribution of sequences within a particular sequence set. It is then possible to apply certain operators that will extract a portion of the set that corresponds to a branch of the taxonomic tree which can then be further analyzed. One has to keep in mind that a set of sequences is not only created when a user uploads one but also when a particular result returns more than one sequence. Alignments from homology searches are treated in a similar manner and the same operators can be applied to them.

4 Conclusions

The large amount of sequence data generated with modern sequencing methods makes the applications that can relate sequences and complex function patterns an absolute necessity. At

the same time, many algorithms for predicting a particular function or uncovering distant evolutionary relationships (which, at the end, allows functional annotations transfer) have become more demanding on compute resources. The output as well as intermediate results can no longer be manually assessed and require sophisticated integrated frameworks. The ANNOTATOR software provides critical support for many protein sequence-analytic tasks by supplying an appropriate infrastructure capable of supporting a large array of sequence-analytic methods, presenting the user with a condensed view of possible functional assignments and, at the same time, allowing to drill down to raw data from the original prediction tool for validation purposes.

References

1. Eisenhaber F (2012) A decade after the first full human genome sequencing: when will we understand our own genome? J Bioinform Comput Biol 10:1271001
2. Kuznetsov V, Lee HK, Maurer-Stroh S, Molnar MJ, Pongor S, Eisenhaber B, Eisenhaber F (2013) How bioinformatics influences health informatics: usage of biomolecular sequences, expression profiles and automated microscopic image analyses for clinical needs and public health. Health Inf Sci Syst 1:2
3. Eisenhaber F, Sung WK, Wong L (2013) The 24th International Conference on Genome Informatics, GIW2013, in Singapore. J Bioinform Comput Biol 11:1302003
4. Pena-Castillo L, Hughes TR (2007) Why are there still over 1000 uncharacterized yeast genes? Genetics 176:7–14
5. Bork P, Dandekar T, az-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y (1998) Predicting function: from genes to genomes and back. J Mol Biol 283:707–725
6. Schneider G, Neuberger G, Wildpaner M, Tian S, Berezovsky I, Eisenhaber F (2006) Application of a sensitive collection heuristic for very large protein families: evolutionary relationship between adipose triglyceride lipase (ATGL) and classic mammalian lipases. BMC Bioinformatics 7:164
7. Eisenhaber F (2006) Bioinformatics: mystery, astrology or service technology. In: Eisenhaber F (ed) Preface for "Discovering Biomolecular Mechanisms with Computational Biology", 1st edn. Landes Biosciences and Eurekah.com, Georgetown, pp 1–10
8. Eisenhaber B, Eisenhaber S, Kwang TY, Gruber G, Eisenhaber F (2014) Transamidase subunit GAA1/GPAA1 is a M28 family metallo-peptide-synthetase that catalyzes the peptide bond formation between the substrate protein's omega-site and the GPI lipid anchor's phosphoethanolamine. Cell Cycle 13:1912–1917
9. Kinoshita T (2014) Enzymatic mechanism of GPI anchor attachment clarified. Cell Cycle 13:1838–1839
10. Novatchkova M, Bachmair A, Eisenhaber B, Eisenhaber F (2005) Proteins with two SUMO-like domains in chromatin-associated complexes: the RENi (Rad60-Esc2-NIP45) family. BMC Bioinformatics 6:22
11. Panizza S, Tanaka T, Hochwagen A, Eisenhaber F, Nasmyth K (2000) Pds5 cooperates with cohesin in maintaining sister chromatid cohesion. Curr Biol 10:1557–1564
12. Prokesch A, Bogner-Strauss JG, Hackl H, Rieder D, Neuhold C, Walenta E, Krogsdam A, Scheideler M, Papak C, Wong WC et al (2011) Arxes: retrotransposed genes required for adipogenesis. Nucleic Acids Res 39:3224–3239
13. Schneider G, Sherman W, Kuchibhatla D, Ooi HS, Sirota FL, Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2012) Protein sequence-structure-function-network links discovered with the ANNOTATOR software suite: application to Elys/Mel-28. In: Trajanoski Z (ed) Computational medicine. Springer, Vienna, pp 111–143
14. Schneider G, Wildpaner M, Sirota FL, Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2010) Integrated tools for biomolecular sequence-based function prediction as exemplified by the ANNOTATOR software environment. Methods Mol Biol 609:257–267
15. Ooi HS, Kwo CY, Wildpaner M, Sirota FL, Eisenhaber B, Maurer-Stroh S, Wong WC,

Schleiffer A, Eisenhaber F, Schneider G (2009) ANNIE: integrated de novo protein sequence annotation. Nucleic Acids Res 37:W435–W440

16. Sherman W, Kuchibhatla D, Limviphuvadh V, Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2015) HPMV: Human protein mutation viewer—relating sequence mutations to protein sequence architecture and function changes. J Bioinform Comput Biol 13 (in press)
17. Eisenhaber F, Bork P (1998) Sequence and structure of proteins. In: Schomburg D (ed) Recombinant proteins, monoclonal antibodies and therapeutic genes. Wiley-VCH, Weinheim, pp 43–86
18. Eisenhaber B, Eisenhaber F, Maurer-Stroh S, Neuberger G (2004) Prediction of sequence signals for lipid post-translational modifications: insights from case studies. Proteomics 4:1614–1625
19. Eisenhaber B, Eisenhaber F (2005) Sequence complexity of proteins and its significance in annotation. In: Subramaniam S (ed) "Bioinformatics" in the encyclopedia of genetics, genomics, proteomics and bioinformatics. Wiley Interscience, New York. doi:10.1002/047001153X.g403313
20. Eisenhaber B, Eisenhaber F (2007) Posttranslational modifications and subcellular localization signals: indicators of sequence regions without inherent 3D structure? Curr Protein Pept Sci 8:197–203
21. Eisenhaber F (2006) Prediction of protein function: two basic concepts and one practical recipe (Chapter 3). In: Eisenhaber F (ed) Discovering biomolecular mechanisms with computational biology, 1st edn. Landes Biosciences and Eurekah.com, Georgetown, pp 39–54
22. Wong WC, Maurer-Stroh S, Eisenhaber F (2010) More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. PLoS Comput Biol 6:e1000867
23. Wong WC, Maurer-Stroh S, Eisenhaber F (2011) Not all transmembrane helices are born equal: towards the extension of the sequence homology concept to membrane proteins. Biol Direct 6:57
24. Sirota FL, Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2015) Single-residue post-translational modification sites at the N-terminus, C-terminus or in-between: to be or not to be exposed for enzyme access. Proteomics 15:2525–2546
25. Eisenhaber F, Wechselberger C, Kreil G (2001) The Brix domain protein family -- a key to the ribosomal biogenesis pathway? Trends Biochem Sci 26:345–347
26. Maurer-Stroh S, Dickens NJ, Hughes-Davies L, Kouzarides T, Eisenhaber F, Ponting CP (2003) The Tudor domain 'Royal Family': Tudor, plant Agenet, Chromo PWWP and MBT domains. Trends Biochem Sci 28:69–74
27. Novatchkova M, Leibbrandt A, Werzowa J, Neubuser A, Eisenhaber F (2003) The STIR-domain superfamily in signal transduction, development and immunity. Trends Biochem Sci 28:226–229
28. Novatchkova M, Eisenhaber F (2004) Linking transcriptional mediators via the GACKIX domain super family. Curr Biol 14:R54–R55
29. Bogner-Strauss JG, Prokesch A, Sanchez-Cabo F, Rieder D, Hackl H, Duszka K, Krogsdam A, Di CB, Walenta E, Klatzer A et al (2010) Reconstruction of gene association network reveals a transmembrane protein required for adipogenesis and targeted by PPARgamma. Cell Mol Life Sci 67:4049–4064
30. Maurer-Stroh S, Ma J, Lee RT, Sirota FL, Eisenhaber F (2009) Mapping the sequence mutations of the 2009 H1N1 influenza A virus neuraminidase relative to drug and antibody binding sites. Biol Direct 4:18
31. Vodermaier HC, Gieffers C, Maurer-Stroh S, Eisenhaber F, Peters JM (2003) TPR subunits of the anaphase-promoting complex mediate binding to the activator protein CDH1. Curr Biol 13:1459–1468
32. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A (2006) Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics Chapter 5, Unit 5.6
33. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A (2007) Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci Chapter 2, Unit 2.9
34. Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9:1753–1773
35. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815
36. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763
37. Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195
38. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, Weese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR et al (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39:D225–D229

39. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15:1000–1011
40. Remmert M, Biegert A, Hauser A, Soding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9:173–175
41. Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244–W248
42. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960
43. Wong WC, Maurer-Stroh S, Eisenhaber F (2011) The Janus-faced E-values of HMMER2: extreme value distribution or logistic function? J Bioinform Comput Biol 9:179–206
44. Wong WC, Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2014) On the necessity of dissecting sequence similarity scores into segment-specific contributions for inferring protein homology, function prediction and annotation. BMC Bioinformatics 15:166
45. Wong WC, Yap CK, Eisenhaber B, Eisenhaber F (2015) dissectHMMER: a HMMER-based score dissection framework that statistically evaluates fold-critical sequence segments for domain fold similarity. Biol Direct 10:39
46. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
47. Wong WC, Maurer-Stroh S, Schneider G, Eisenhaber F (2012) Transmembrane helix: simple or complex. Nucleic Acids Res 40:W370–W375
48. Kreil DP, Ouzounis CA (2003) Comparison of sequence masking algorithms and the detection of biased protein sequence regions. Bioinformatics 19:1672–1681
49. Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. Bioinformatics 16:915–922
50. Iakoucheva LM, Dunker AK (2003) Order, disorder, and flexibility: prediction from protein sequence. Structure 11:1316–1317
51. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003) Protein disorder prediction: implications for structural proteomics. Structure 11:1453–1459
52. Linding R, Russell RB, Neduva V, Gibson TJ (2003) GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res 31:3701–3708
53. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433–3434
54. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347:827–839
55. Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S (1992) Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci U S A 89:2002–2006
56. Claverie JM (1994) Large scale sequence analysis. In: Adams MD, Fields C, Venter JC (eds.), Automated DNA sequencing and analysis. Academic Press, San Diego, pp. 267–279.
57. Claverie JM, States DJ (1993) Information enhancement methods for large scale sequence analysis. Comput Chem 17:191–201
58. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337:635–645
59. Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem 17:149–163
60. Wootton JC (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 18:269–285
61. Wootton JC (1994) Sequences with "unusual" amino acid compositions. Curr Opin Struct Biol 4:413–421
62. Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266:554–571
63. Eisenhaber B, Bork P, Eisenhaber F (1999) Prediction of potential GPI-modification sites in proprotein sequences. J Mol Biol 292:741–758
64. Eisenhaber B, Wildpaner M, Schultz CJ, Borner GH, Dupree P, Eisenhaber F (2003) Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for Arabidopsis and rice. Plant Physiol 133:1691–1701
65. Eisenhaber B, Maurer-Stroh S, Novatchkova M, Schneider G, Eisenhaber F (2003) Enzymes and auxiliary factors for GPI lipid anchor biosynthesis and post-translational transfer to proteins. Bioessays 25:367–385

66. Eisenhaber B, Schneider G, Wildpaner M, Eisenhaber F (2004) A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for Aspergillus nidulans, Candida albicans, Neurospora crassa, Saccharomyces cerevisiae and Schizosaccharomyces pombe. J Mol Biol 337:243–253
67. Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2002) N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. J Mol Biol 317:541–557
68. Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2002) N-terminal N-myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences. J Mol Biol 317:523–540
69. Maurer-Stroh S, Gouda M, Novatchkova M, Schleiffer A, Schneider G, Sirota FL, Wildpaner M, Hayashi N, Eisenhaber F (2004) MYRbase: analysis of genome-wide glycine myristoylation enlarges the functional spectrum of eukaryotic myristoylated proteins. Genome Biol 5:R21
70. Maurer-Stroh S, Eisenhaber F (2004) Myristoylation of viral and bacterial proteins. Trends Microbiol 12:178–185
71. Maurer-Stroh S, Washietl S, Eisenhaber F (2003) Protein prenyltransferases. Genome Biol 4:212
72. Maurer-Stroh S, Eisenhaber F (2005) Refinement and prediction of protein prenylation motifs. Genome Biol 6:R55
73. Maurer-Stroh S, Koranda M, Benetka W, Schneider G, Sirota FL, Eisenhaber F (2007) Towards complete sets of farnesylated and geranylgeranylated proteins. PLoS Comput Biol 3, e66
74. Neuberger G, Maurer-Stroh S, Eisenhaber B, Hartig A, Eisenhaber F (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. J Mol Biol 328:581–592
75. Neuberger G, Maurer-Stroh S, Eisenhaber B, Hartig A, Eisenhaber F (2003) Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. J Mol Biol 328:567–579
76. von Heijne G (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 14:4683–4690
77. von Heijne G (1987) Sequence analysis in molecular biology? Treasure trove or trivial pursuit. Academic, San Diego
78. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795
79. Nielsen H, Engelbrecht J, Brunak S, von HG (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10:1–6
80. Nielsen H, Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. Proc Int Conf Intell Syst Mol Biol 6:122–130
81. Cserzo M, Eisenhaber F, Eisenhaber B, Simon I (2002) On filtering false positive transmembrane protein predictions. Protein Eng 15:745–752
82. Cserzo M, Eisenhaber F, Eisenhaber B, Simon I (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics 20:136–137
83. Tusnady GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283:489–506
84. Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036
85. Kall L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res 35:W429–W432
86. Krogh A, Larsson B, von HG, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580
87. Sonnhammer EL, Von HG, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6:175–182
88. Claros MG, von Heijne G (1994) TopPred II: an improved software for membrane protein structure predictions. Comput Appl Biosci 10:685–686
89. von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225:487–494
90. Lupas A, Van DM, Stock J (1991) Predicting coiled coils from protein sequences. Science 252:1162–1164
91. Lupas A (1996) Prediction and analysis of coiled-coil structures. Methods Enzymol 266:513–525
92. Frishman D, Argos P (1996) Incorporation of non-local interactions in protein secondary

structure prediction from the amino acid sequence. Protein Eng 9:133–142

93. Frishman D, Argos P (1997) Seventy-five percent accuracy in protein secondary structure prediction. Proteins 27:329–335
94. Eisenhaber F, Imperiale F, Argos P, Frommel C (1996) Prediction of secondary structural content of proteins from their amino acid composition alone. I New analytic vector decomposition methods. Proteins 25:157–168
95. Eisenhaber F, Frommel C, Argos P (1996) Prediction of secondary structural content of proteins from their amino acid composition alone. II The paradox with secondary structural class. Proteins 25:169–179
96. Maurer-Stroh S, Gao H, Han H, Baeten L, Schymkowitz J, Rousseau F, Zhang L, Eisenhaber F (2013) Motif discovery with data mining in 3D protein structure databases: discovery, validation and prediction of the U-shape zinc binding ("Huf-Zinc") motif. J Bioinform Comput Biol 11:1340008
97. Andrade MA, Ponting CP, Gibson TJ, Bork P (2000) Homology-based method for identification of protein repeats using statistical significance estimates. J Mol Biol 298:521–537
98. Andrade MA, Petosa C, O'Donoghue SI, Muller CW, Bork P (2001) Comparison of ARM and HEAT protein repeats. J Mol Biol 309:1–18
99. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39:W339–W346
100. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, Weber T (2013) antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res 41:W204–W212
101. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Muller R, Wohlleben W et al (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43:W237–W243
102. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40:W445–W451
103. Desai DK, Nandi S, Srivastava PK, Lynn AM (2011) ModEnzA: accurate identification of metabolic enzymes using function specific profile HMMs with optimised discrimination threshold and modified emission probabilities. Adv Bioinformatics 2011:743782
104. Wolf YI, Brenner SE, Bash PA, Koonin EV (1999) Distribution of protein folds in the three superkingdoms of life. Genome Res 9:17–26
105. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 3:265–274
106. Sigrist CJ, de CE, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41:D344–D347
107. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A et al (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31:3625–3630
108. Berezovsky IN, Grosberg AY, Trifonov EN (2000) Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett 466:283–286
109. Goncearenco A, Berezovsky IN (2010) Prototypes of elementary functional loops unravel evolutionary connections between protein functions. Bioinformatics 26:i497–i503
110. Goncearenco A, Berezovsky IN (2015) Protein function from its emergence to diversity in contemporary proteins. Phys Biol 12:045002
111. Mott R (2000) Accurate formula for P-values of gapped local sequence and profile alignments. J Mol Biol 300:649–659
112. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
113. Dayhoff M (1979) Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, DC
114. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C (2011) OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res 39:D289–D294
115. Roth AC, Gonnet GH, Dessimoz C (2008) Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics 9:518
116. Biegert A, Soding J (2009) Sequence context-specific profiles for homology searching. Proc Natl Acad Sci U S A 106:3770–3775
117. Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. J Mol Biol 276:71–84

118. Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132:185–219
119. Sirota FL, Ooi HS, Gattermayer T, Schneider G, Eisenhaber F, Maurer-Stroh S (2010) Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. BMC Genomics 11(Suppl 1):S15
120. Enright AJ, Van DS, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584
121. van Dongen S (2008) Graph clustering via a discrete uncoupling process. SIAM J Matrix Anal Appl 30:121–141
122. Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17:282–283
123. Li W, Jaroszewski L, Godzik A (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics 18:77–82
124. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659
125. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217
126. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797
127. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113
128. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res 15:330–340
129. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066
130. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518
131. Katoh K, Toh H (2007) PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. Bioinformatics 23:372–374
132. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286–298