Jakub Paś

Application and implementation of probabilistic profile-profile comparison methods for protein fold recognition

(pol.: Wdrożenie i zastosowania probabilistycznych metod porównawczych profil-profil w rozpoznawaniu pofałdowania białek)

Rozprawa doktorska
ma formę spójnego tematycznie zbioru artykułów
opublikowanych w czasopismach naukowych*

Wydział Chemii, Uniwersytet im. Adama Mickiewicza w Poznaniu
Promotor: dr hab. Marcin Hoffmann, prof. UAM
Promotor pomocniczy: dr Krystian Eitner

* Ustawa o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki

Dz.U.2003.65.595 - Ustawa z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki

Art 13

1. Rozprawa doktorska, przygotowywana pod opieką promotora albo pod opieką promotora i promotora pomocniczego, o którym mowa w art. 20 ust. 7, powinna stanowić oryginalne rozwiązanie problemu naukowego lub oryginalne dokonanie artystyczne oraz wykazywać ogólną wiedzę teoretyczną kandydata w danej dyscyplinie naukowej lub artystycznej oraz umiejętność samodzielnego prowadzenia pracy naukowej lub artystycznej.

2. Rozprawa doktorska może mieć formę maszynopisu książki, książki wydanej lub spójnego tematycznie zbioru rozdziałów w książkach wydanych, spójnego tematycznie zbioru

artykułów opublikowanych lub przyjętych do druku w czasopismach naukowych, określonych przez ministra właściwego do spraw nauki na podstawie przepisów dotyczących finansowania nauki, jeżeli odpowiada warunkom określonym w ust. 1.

3. Rozprawę doktorską może stanowić praca projektowa, konstrukcyjna, technologiczna lub artystyczna, jeżeli odpowiada warunkom określonym w ust. 1.

4. Rozprawę doktorską może także stanowić samodzielna i wyodrębniona część pracy zbiorowej, jeżeli wykazuje ona indywidualny wkład kandydata przy opracowywaniu koncepcji,

wykonywaniu części eksperymentalnej, opracowaniu i interpretacji wyników tej pracy, odpowiadający warunkom określonym w ust. 1.

5. Za zgodą rady jednostki przeprowadzającej przewód, rozprawa doktorska może być przedstawiona w języku innym niż polski.

6. Rozprawa doktorska powinna być opatrzona streszczeniem w języku angielskim, a rozprawa doktorska przygotowana w języku obcym również streszczeniem w języku polskim. W przypadkach, gdy rozprawa doktorska nie ma formy pisemnej powinna być opatrzona opisem w języku polskim i angielskim.

7. Streszczenie rozprawy doktorskiej łącznie z recenzjami zamieszcza się na stronie internetowej szkoły wyższej lub jednostki organizacyjnej przeprowadzającej przewód doktorski.

Streszczenie rozprawy doktorskiej zamieszcza się w dniu podjęcia przez radę jednostki uchwały o przyjęciu rozprawy doktorskiej, a recenzje w dniu ich przekazania przez recenzentów. Streszczenie rozprawy i recenzje pozostają na stronie internetowej co najmniej do dnia nadania stopnia doktora. Warunek zamieszczenia streszczenia rozprawy doktorskiej i recenzji nie dotyczy rozprawy doktorskiej, której przedmiot jest objęty ochroną informacji niejawnych.

8. Recenzje podlegające zamieszczeniu na stronie internetowej przekazuje się niezwłocznie po ich złożeniu do Centralnej Komisji w celu ich opublikowania w Biuletynie Informacji Publicznej.

Lista artykułów opublikowanych w czasopismach naukowych:

1. ELM: the status of the 2010 eukaryotic linear motif resource
   CM Gould, F Diella, A Via, P Puntervoll, C Gemünd, S Chabanis-Davidson, ... **J Pas**, ...
   Nucleic acids research 38 (suppl 1), D167-D180 139 2010

Mój wkład przy opracowywaniu serwisu PDB BLASprzewidującego strukturę białek na podstawie sekwencji oraz opracowanie algorytmu do oznaczania zachowawczych motywów aminokwasowych oraz opracowywaniu wyników granty ELM.

2. ORFeus: detection of distant homology using sequence profiles and predicted secondary structure
   K Ginalski, **J Pas**, LS Wyrwicz, M Von Grotthuss, JM Bujnicki, L Rychlewski
   Nucleic acids research 31 (13), 3804-3807 126 2003

Wkład mój dotyczył przygotowania baz dancyh algorytmu, benchmarków opracowywania wyników oraz przygotowania manuskryptu.

3. Molecular phylogenetics of the RrmJ/fibrillarin superfamily of ribose 2'-O-methyltransferases
   M Feder, **J Pas**, LS Wyrwicz, JM Bujnicki
   Gene 302 (1-2), 129-138 66 2003

Wkład mój dotyczył wyszukiwania homologów superrodziny RrmJ/fibrillarin oraz wykonania modeli 3D białek z tej rodziny.

4. Ligand. info small-molecule meta-database
   M Grotthuss, G Koczyk, **J Pas**, LS Wyrwicz, L Rychlewski
   Combinatorial Chemistry & High Throughput Screening 7 (8), 757-761 46 2004

Wkład mój dotyczył przygotowania infrastruktury bazy danych, opracowania wyników oraz manuskryptu.

5. Application of 3D Jury, GRDB, and Verify3D in fold recognition
   M Grotthuss, **J Pas**, L Wyrwicz, K Ginalski, L Rychlewski
   Proteins: Structure, Function, and Bioinformatics 53 (S6), 418-423 45 2003

Wkład mój dotyczył udziału w eksperymencue CASP, wyszukiwaniu sekwencji homologicznych, przygotowywaniu modeli 3D białek, przygotowania baz dancyh algorytmu, benchmarków opracowywania wyników oraz przygotowania manuskryptu.

6. Ligand-Info, searching for similar small compounds using index profiles
   M von Grotthuss, **J Pas**, L Rychlewski
   Bioinformatics 19 (8), 1041-1042 38 2003

Wkład mój dotyczył przygotowania infrastruktury bazy danych, opracowania wyników oraz manuskryptu.

7. Analysis of structure and function of tenascin-C
   J Pas, E Wyszko, K Rolle, L Rychlewski, S Nowak, R Żukiel, J Barciszewski
   The international journal of biochemistry & cell biology 38 (9), 1594-1602 34 2006

Wkład mój dotyczył koncepcji, przewidywania elementów strukturalnych tenascyny-C, przewidywania struktury trzeciorzędowej, analizy motywów strukturalnych, opracowania wyników i przygotowania manuskryptu.

8. Structure prediction, evolution and ligand interaction of CHASE domain
   J Pas, M von Grotthuss, LS Wyrwicz, L Rychlewski, J Barciszewski
   FEBS letters 576 (3), 287-290 26 2004

Wkład mój dotyczył koncepcji, analizy filogenetycznej, analizy topologicznej białka, przewidywania struktury trzeciorzędowej, dokowania molekularnego, opracowania wyników i przygotowania manuskryptu.

9. 3D-Hit: fast structural comparison of proteins.
   D Plewczyński, **J Pas**, M Von Grotthuss, L Rychlewski
   Applied bioinformatics 1 (4), 223 25 2002

Wkład mój dotyczył przygotowania infrastruktury bazy danych, opracowania wyników oraz manuskryptu.

10. Lead toxicity through the leadzyme
    MZ Barciszewska, M Szymanski, E Wyszko, **J Pas**, L Rychlewski, J Barciszewski
    Mutation Research/Reviews in Mutation Research 589 (2), 103-110 18 2005

Wkład mój opracowaniu i interpretacji wyników tej pracy oraz przygotowania manuskryptu.

11. Comparison of proteins based on segments structural similarity
    D Plewczynski, **J Pas**, M von Grotthuss, L Rychlewski
    Acta Biochimica Polonica 51 (1), 161-172 11 2004

Wkład mój dotyczył przygotowania infrastruktury bazy danych, opracowania wyników oraz manuskryptu.

12. How unique is the rice transcriptome?
    LS Wyrwicz, M von Grotthuss, **J Pas**, L Rychlewski
    Science (New York, NY) 303 (5655), 168; author reply 168 5 2004

Wkład mój dotyczył przygotowania infrastruktury, bazy danych, opracowania wyników oraz manuskryptu.

13. Two sequences encoding chalcone synthase in yellow lupin (Lupinus luteus l.) may have evolved by gene duplication
    D Narożna, **J Pas**, J Schneider, CJ Mądrzak
    Cellular & molecular biology letters 9 (1), 95-105 5 2004

Wkład mój dotyczył koncepcji, analizy filogenetycznej, modelowania molekularnego, opracowania wyników oraz manuskryptu.

14. Predicting protein structures accurately
    L Rychlewski, LS Wyrwicz, M Von Grotthuss, **J Pas**
    Science 304 (5677), 1597 5 2004

Wkład mój dotyczył opracowywania wyników oraz przygotowania manuskryptu.

15. The PDB-Preview database: a repository of in-silico models of ‘on-hold’PDB entries
    D Fischer, **J Pas**, L Rychlewski
    Bioinformatics 20 (15), 2482-2484 3 2004

Wkład mój dotyczył koncepcji, przygotowania infrastruktury, bazy danych, opracowania wyników oraz manuskryptu.

16. Leadzyme formed in vivo interferes with tobacco mosaic virus infection in Nicotiana tabacum
    E Wyszko, M Nowak, H Pospieszny, M Szymanski, **J Pas**, MZ Barciszewska, J ...
    FEBS Journal 273 (22), 5022-5031 2 2006

Wkład mój dotczył wykonywania części eksperymentalnej, opracowaniu i interpretacji wyników tej pracy oraz przygotowania manuskryptu.

17. GRDB - Gene Relational DataBase
    J Pas, P Stępniak, L Wyrwicz, K Ginalski, L Rychlewski
    BioInfoBank Library Acta 11 (1), 2659 2011

Wkład mój dotyczył przygotowania baz dancyh, infrastruktury, benchmarków, opracowywania wyników oraz przygotowania manuskryptu.

Jakub Paś

Streszczenie w języku polskim

Jakub Paś

Wdrożenie i zastosowania probabilistycznych metod porównawczych profil-profil w rozpoznawaniu pofałdowania białek

Metody rozpoznawania pofałdowania białka zwane też rozpoznawaniem foldów (eng. Fold Recognition) są metodami wykrywania i przewidywania struktury trzeciorzędowej białka, stosowanymi dla białek, które nie posiadają sekwencji homologicznych o znanej strukturze trzeciorzędowej, zdeponowanych w międzynarodowej bazie danych struktur białkowych (eng. Protein Data Bank). Metody te opierają się na założeniu, że w wyniku ewolucji oraz ograniczeń fizycznych i chemicznych w przyrodzie znajduje się określona i ograniczona liczba odmiennych zwojów białek

Metody Rozpoznawania ufałdowania białka są wykorzystywane do przewidywania struktury białek, analizy ewolucyjnej, analizy szlaków metabolicznych, enzymatycznych, przewidywania skuteczności dokowania molekularnego i projektowania leków.

Obecnie istnieje około 1300 odkrytych i scharakteryzowanych foldów białek zgrupowanych w bazach danych takich jak SCOP czy CATCH. Każde nowo odkryte białko ma duże szanse by zostać sklsyfikowane jako członek jednej z takich grup. Dotychczas zostało zaproponowanych wiele odmiennych podejść w znajdowaniu poprawnego foldu dla nowo scharakteryzowanych sekwencji. Zwykle wykorzystuje się do tego informacje o ewolucji zarówno sekwencji poszukiwanej jak i sekwencji docelowych. Jedną z metod wykorzystujących takie informacje to porównanie profilu białkowego sekwencji poszukiwanej z profilami sekwencji w bazach dancych zawierajcacych znane struktury. Metody takie nazywane są metodami porównywania typu profil-profil.

Uliniowienia w profilach sekwencyjnych metod profil-profil mogą być obliczane przy pomocy iloczynu skalarnego, modelu probablistycznego, stochastycznego albo przy pomocy miar teoretycznych. Zaprezentowane tu zastosowania i wdrożenia metod porównywania białek typu profil-profil wskazują na zalety zastosowania probablistycznych funkcji oceniających jakość porównania profili nad innymi metodami rozpoznawania foldów.

Celem pracy jest wskazanie iż metody porównywania profil-profil mogą przewyższać inne metody rozpoznawania foldów w analizie spokrewnionych białek, i że mogą być one stosowane nie tylko do rozpoznawania foldów, ale także do innych celów takich jak wykrywanie i identyfikacja genów, granic domen białkowych oraz modelowania złożonych struktur białkowych.

Jakub Paś

Application and implementation of probabilistic profile-profile comparison methods for protein fold recognition

I. Summary

Fold recognition is a method of fold detecting and protein tertiary structure prediction applied for proteins lacking homologues sequences of known fold and structure deposited in the Protein Data Bank. They are based on assumption that there is strictly limited number of different protein folds in nature, mostly as a result of evolution and due to basic physical and chemical constraints of polypeptide chains.
Fold recognition methods are useful for protein structure prediction, evolutionary analysis, metabolic pathways and enzymatic efficiency prediction, molecular docking and drug design.
Currently there are about 1300 discovered and characterized protein folds in SCOP and CATH databases. Every newly discovered protein sequence has significant chances to be classified into one of those folds. Many different approaches have been proposed for finding the correct fold for a new sequence and it is often useful to include evolutionary information for query as well as for target proteins. One of the methods of including this information is a comparison of a query and target sequences profiles. These fold recognition techniques are called profile-profile methods.
Profile-profile alignments can be calculated using a dot-product, a probabilistic model, stochastic or theoretical measures. Here are presented applications and implementations of probabilistic profile-profile comparison methods and advantages of usage of probabilistic scoring function over comparable fold recognition techniques.
The purpose of this comparison is to show that probabilistic profile-profile methods may outperform other fold recognition methods in comparison in analysis of distantly related proteins and that they can be applied not only for fold recognition but also for slightly different purposes like gene identification[1], detection of domain boundaries and modeling of complex proteins[2].

II. Introduction

Since the insulin protein was characterized by Fred Sanger in 1951[3] millions of protein sequences have been identified. The evolutionary relationships between these proteins can be discovered by aligning them together to show their similarities and assigning such alignment numerical values defines their evolutionary distance. There are two main types of sequence alignment. The pair wise alignment which compares two different sequences and multiple sequence alignment in which many sequences are compared.

Chronologically the first and most frequently used algorithms for pairwise sequence comparison are Needleman-Wunsch[4], the Smith-Waterman and BLAST[5]. Most common application for pairwise sequence alignment is comparison of two sequences under study and database searching for homologous sequences which usually means performing a of comparison of sequence in study with all know sequences stored in database. Compared sequences are generally ordered by similarity to show the most closely related sequences and the matches are usually reported with a measure of statistical significance.

For multiple sequence alignment the ClustalW[6] and T-Coffee[7] are most popular algorithms. Multiple alignments are usually used to detect relationships within the group of similar sequences like protein family to show evolutionary relationships. As the multiple sequences alignment contains more information then pair wise alignment the multiple sequence alignment can be used for a more sensitive study of two sequences comparison by using evolutionary information from neighboring sequences.

The problem arises when the two sequences are so distant that simple pairwise alignment cannot be used to comparison.

1. Sequence comparison methods used for fold recognition.

Before the profile-profile methods were discovered the simple sequence-sequence methods were used for fold recognition. Those methods can be divided to global and local aligning methods. In principle the global alignment is the comparison of entire two sequences regardless that some parts of those sequences can be aligned poorly or the reasonable alignment can not be performed at all. The local alignment performs only the alignment of one or more conserved parts of the sequences in study excluding the not conserved parts from comparison.

Historically the basic algorithm for computing an optimal alignment of two sequences was independently developed by different scholars from different scientific disciplines:

Vintsyuk in 1968[8] for speech processing, Needleman-Wunsch in 1970[4] for molecular biology and Wagner-Fischer in 1974 [9] for computer science. Basic principle of those algorithms is the same and only the Needleman & Wunsch are cited in biomedical sciences.

1.1. Needleman-Wunsch

Needleman and Wunsch designed their algorithm explicitly for the case when the alignment is penalized by the matches and mismatches and gaps representing insertions and deletions have no penalty. An algorithm performs the optimal global pair wise alignment of two sequences using dynamic programming. The basic idea was to build an optimal alignment using optimal alignments of smaller subsequences[10].
To perform this task all segments of the sequences are compared with each other. The algorithm recursively calculates the total scores for all subsequences from top to bottom and from left to right. In each recursion step, a specific scoring function is employed in order to evaluate matching segments. In order to obtain the alignment of the sequences a trace back function had to be applied to find the path of best choices. The scoring function depends on the purpose of the analysis. The simplest way is to penalize mismatches and matches with 1 and 0 respectively. The matches means the exact, identical amino acid or amino acid with similar biochemical computed from substitution matrix like log-odds matrices like PAM (Point Accepted Mutation) [11] or BLOSUM (BLOck SUbstitution Matrix)[12]. The original publication[4] suggests usage of the recursion and arbitrary gap for penalization. More efficient version of this algorithm was introduced by Sankoff in 1972 [13].
Needleman and Wunsch formulated their problem in terms of maximizing similarity but other approach could be to minimize distance between sequences which is an equivalent computational problem. In modern terminology, "Needleman-Wunsch" refers to the global alignment similarly like “Smith-Waterman” refers to local alignment.

1.2. Smith-Waterman

The Smith-Waterman algorithm is a well known algorithm for performing local sequence alignment between two sequences instead of looking at the total sequence lenght. The algorithm was first proposed by Temple F. Smith and Michael S. Waterman in 1981[14] and it is a variation of Needleman-Wunsch algorithm. The main difference to the Needleman-Wunsch is that negative scoring matrix cells are set to zero so only local alignments are visible. Smith-Waterman is also dynamic programming algorithm which means that is guaranteed to find the optimal local alignment with respect to the scoring system being used. Backtracking starts at the highest scoring matrix cell and proceeds

until a cell with score zero is encountered providing the highest scoring local alignment. Currently multiple improved alternatives of Smith-Waterman are available[15, 16]. The Smith-Waterman algorithm is time consuming for large sequences. A linear time algorithms such BLAST has been modified to decrease the amount of time required to identify conserved regions between two sequences under study.

1.3. BLAST

BLAST (Basic Local Alignment Search Tool) is the algorithm for comparison of biological sequences designed to perform fast database searches. BLAST enables a researcher to compare a query sequence with a library or database of sequences with similarity above a certain threshold in reasonable time. The BLAST program was designed by Stephen Altschul, and David J. Lipman in 1990[5]. It is one of the most widely used bioinformatic heuristic algorithms and it is much faster than calculating an optimal alignment using Smith-Waterman. This emphasis on speed is crucial to make the algorithm practical in searching against the huge genome databases. While BLAST is faster than Smith-Waterman, it cannot guarantee the optimal alignments of the query and database sequences. Because of its speed the BLAST algorithm is widely used for creating sequence profiles by iteratively performing searches with the results of previous queries. Such implementation of BLAST is used in PSI-BLAST program.

2. Sequence-Profile methods.

In principle the sequence-profile fold recognition methods use the initially build profiles which are then compared with sequences. In 1987 Michael Gribskov introduced the method of aligning two sequences by comparison of multiple sequence alignment of one sequence to another sequence [17]. The representation of such multiple alignment is also known as Position Specific Scoring Matrices (PSSM) which is commonly used for representation of sequences, motifs and patterns in biological data. PSSM is a matrix of score values that gives a weighted match to any given string of fixed length. It has one row for each symbol of the alphabet (for every amino acid or DNA base) and one column for each position in the pattern. The probabilistic interpretation of profiles was introduced by Brown and Haussler in 1993[18] using hidden Markov models. These models have become known as HMM - profiles. In sequence – profile methods the profiles can by build for both query or target sequence.

2.1. PSI - BLAST

PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) was historically the first algorithm that performs sequence-profile comparison[19]. PSI-

BLAST builds a profile which is practically a PSSM from sequences aligned witch the score below given threshold. Such profile is then used to further search the database for new matches.
The PSI-BLAST program is often much more sensitive than the corresponding BLAST program but it takes a little more time to perform comparison because of multiple iterations during search. In work of Henikoff and Henikoff [20] it was described how to modify BLAST so that it may operate on a position-specific score matrix- a single virtual sequence that may be used as a query with the original BLAST program. The alignment of position-specific score matrix to sequence is almost completely analogous to the alignment of two sequences. The only real difference is that the score for aligning a amino acid letter with a pattern from position is taken directly form PSSM under study rather than from default substitution matrix. Position-specific protein score matrices draw their power from its improved estimation of the probabilities of amino at given positions and from relatively precise definition of the boundaries of important motifs. The greatest challenge of automatic process off building PSSM is to set the universal constraints which may be used for searching against distinct protein families. Also the query sequence may contain multiple different protein domains so the preparation of query sequence and interpretation of the results may be challenging and may require human supervision.

2.2. RPS-BLAST

RPS-BLAST is also known as Reverse PSI-BLAST. It searches a query sequence against a database of profiles[21]. In principle this is the opposite of PSI-BLAST which searches a profile against a database of sequences. RPS-BLAST uses a BLAST like algorithm, finding single or double word hits and then performs an ungapped extension on this candidate. If a sufficiently high-scoring ungapped alignment is produced, a gapped extension is performed and those gapped alignments with sufficiently low expect value are reported. The similar attempt was used in IMPALA[22] method which instead of BLAST performs a Smith-Waterman calculation between the query and each profile which is more time consuming. RPS-BLAST uses a BLAST and pre-computed profiles to allow the search to proceed faster. The natural evolution of methods such as PSI-BLAST, RPS-BLAST or IMPALA are the profile–profile methods where for both query and target sequences the profiles needs to be pre-computed before database searches can be performed.

3. Profile-profile methods.

In recent years numerous methods have been developed that allow direct comparison of profiles[23]. Profile-profile alignments can be implemented in several fundamentally

different ways. The similarity between two positions can be calculated mainly by three different methods: using a dot-product, probabilistic model or an information theoretical measure.
Profile-profile methods perform much better than standard sequence-profile methods both in their ability to recognize superfamily-related proteins and in the quality of the obtained alignments. Although the performance between different profile-profile methods is quite similar, methods using probabilistic scoring function have an advantage as they can create good alignments and show a good fold recognition capacity using the same gap-penalties, while the other methods need to use several different parameters at the same time to obtain comparable performances[23].

The fundamental difference between different profle-profle alignment methods lies in how they calculate the score between two profile positions. A profile or PSSM is simply a set of vectors, where each vector contains the frequency of each type of amino acid in a particular position of the multiple sequence alignment. In sequence-profile alignments like PSI-BLAST[19], RPS-BLAST[21] or IMPALA[22] the score is calculated by extracting logarithm of the probability for an amino acid in this vector. However, in profile-profile alignments, we have to compare two frequency vectors and this can be done in several different ways, including calculating the sum of pairs, the dot-product, or a correlation coefficient between the two vectors. In addition, information about the background frequency can be used. The performance of profile-profile methods depends on: calculation of the score between two profile positions, alignment methodology and score function to measure evolutionary distance between two sequences in study
Profile–profile methods are better at detecting distantly related proteins and provide better alignments for these proteins than sequence–sequence, and sequence-profile methods. Profile-profile scoring methods are better at distinguishing evolutionary related positions from non-related positions.
In this study the implementations and applications of four different profile-profile methods are described: FFAS[24], Meta-BASIC[25] and ORFEUS[26].

3.1. BASIC (Bilateral Amplified Sequence Information Comparison)

The BASIC [27] algorithm was historically first algorithm which utilized procedure for profile calculation and using profiles on both sides of the alignment. This algorithm compares a sequence profile of a query protein to a library of profiles representing known protein structures. The two main differences between this algorithm and PDB-BLAST[24] algorithm is a different, simplified procedure for profile calculation which is applied for query and each target profile. Two sequence profiles are compared using Smith-Waterman dynamic programming algorithm. The similarity score between

positions in two sequences is calculated with the mutation matrix such as for the Gonnett similarity matrix [28]. For two profiles, this value is calculated as an average of scores between all amino acid pairs, averaged according to the probability distribution in each profile. Three parameters, gap introduction penalty, gap extension penalty and a constant, added to each element of the mutation matrix are optimized for a fold recognition benchmark.

3.2. FFAS (Fold & Function Assignment)

This algorithm is similar to BASIC and can be recognized as it's improved version. FFAS similarly as BASIC uses profile information on both sides of the alignment, but it is based on a novel procedure for profile preparation from the multiple alignments of sequences in the family of homologous proteins[29].
Calculation of a multiple sequence alignment for each profile is performed by five iterations of PSI-BLAST against the sequence pool database nr85s (Non Redundand GenBank[30] protein database clustered as 85% of identity). Then calculation of a sequence profile using sequences found by PSI-BLAST is performed. Weights are assigned to sequences based on their uniqueness.
FFAS aligns profiles using a standard local-local dynamic programming algorithm. The value of the comparison score between positions n and m from the two profiles is calculated as a vector*matrix*vector product that includes the n-th column from the first profile, the substitution matrix BLOSUM62 [20], and the m-th column from the second profile. The alignment score is then calculated using dynamic programming.
Finally the calculation of the final FFAS score is performed by comparing it with the distribution of scores obtained for pairs of unrelated proteins.
FFAS algorithm well balances two effects: introducing of new information, which leads to increased sensitivity in recognizing distant homologues with avoidance of errors, leading to incorrect homology assignments.

3.3. ORFEUS

ORFeus is a fully automated, sensitive protein sequence similarity search algorithm available to the academic community via the Structure Prediction Meta Server[31]. The goal of the method was to increase the sensitivity of the detection of distantly related protein families by adding secondary structure information. The technique which combines the information from analysis of protein structures with information coming from the analysis of homologous proteins families is also called "hybrid threading" approach[32]. The alignment of meta-profiles created this way is more sensitive in detecting remote homology between protein families. The specificity of the alignment

score is improved in the lower specificity range compared with the sequence-only profiles[26].

The secondary structure prediction is stored in the form of a profile of probabilities. ORFeus can utilize any secondary structure prediction method that produces estimated probabilities for local structure described using three states, that is, the helix, the beta sheet and the loop produced by PSIPRED [33] method. The sequence profiles are generated as in FFAS [29]. The main difference is that all the vectors of probabilities for the occurrence of all amino acids at each position are normalized. The similarity between two positions equals sum of the shifted dot product of the sequence profile and the shifted dot product of the secondary structure probability vector multiplied by the secondary structure weight at the given position.

The combined local alignment of two sequence profiles and two secondary structure profiles conducted by ORFeus requires five parameters: gap initiation penalty, gap extension penalty, a weight for the contribution of the secondary structure profiles and two values, which shift the expected dot product of the secondary structure and sequence vectors below zero (expected score of aligning two vectors representing two residues). All five parameters were selected using brute-force optimization on a test set of artificially constructed two-domain families. A genetic algorithm was used to evolve and improve the program parameters.
Two types of scoring functions were used for the optimization of parameters. The total sensitivity score for the test set was measured as the sum of prediction scores over all 118 targets. Each prediction score, calculated for each target, is the sum of all correct hits scaled by the number of wrong hits with higher alignment score.

Only one top-scoring prediction for each family is taken into account. This corresponds to the common procedure of specificity evaluation conducted in the LiveBench [34]. The performance of optimized ORFeus algorithm was compared with version without secondary structure and with PSI-Blast. The results show that the more complex meta profiles that utilize predicted secondary structure preferences are 10% more specific than the simple sequence-only profiles. The sensitivity of the meta profiles conducted by ORFeus is able to boost the sensitivity even further, providing up to 50% improvement compared with other two methods.

4. Other fold recognition methods

4.1. Threading

The word threading implies that the protein query sequence is dragged step by step through each possible position on each template to search for the best arrangement of the sequence as measured by score function which is usually quasi-energy function.
The name for prediction method[35] came from observation that one can verify the predicted protein structure using atomic representation of the protein template[36].
Finding the best arrangement of residues, including gaps and insertions is the problem of sequence to structure alignment. Because threading calculations may by computation time consuming the initial library of target structure is usually reduced by initial homology search. Protein threading treats the template in an alignment as a structure, and both sequence and structure information extracted from the alignment are used for prediction. When there is no significant homology found, protein threading can make a prediction based on the structure information. That also explains why protein threading may be more effective than homology modeling in many cases. In practice, when the sequence identity in a sequence alignment is low homology modeling may not produce a significant prediction. In this case, if there is distant homology found for the target, protein threading can generate a prediction. The closer a template is to the correct answer; the more likely the sequence is to score well on it. Statistically, the well represented fold is more likely to score well by chance. The threading methods are often use for fold recognition and the final alignment is refined by other methods.
Threading score functions are usually more simplified than those used in a real energy calculation. In a threading calculation, the sequence residues are placed on the backbone of the template structure and from there, one can calculate ideal coordinates for the C-beta atom. Consequently, a threading score function usually represents each residue by one or a few interaction sites.
Most popular threading software are: HHpred[37], RaptorX[38], MUSTER[39] and relatively new method SPARKS-X[40].

4.2. "Ab Initio" protein structure prediction methods

"Ab initio" methods are "holy grail" of the fold recognition and protein structure prediction methods. The aim of such methods is to provide protein models built "*de novo*" based on bio-physical principles rather than directly on previously solved structures. There are multiple possible procedures performing *"ab initio"* prediction. Some of them perform simulated folding; the same way at it is performed in cell during translation process. Other uses the stochastic method to search possible conformations of protein by global

optimization of a suitable energy function. All procedures tend to require significant amount of computational power to overcome so called “Levinthal's paradox”. As it was observed by Cyrus Levinthal[41] small protein of 100 residues may misfold up to 3^{198} conformations (where 198 is the number of possible different phi and psi bond angles). It is not possible to check all those conformations in reasonable time even though in the cell folding process takes miliseconds. Also in many cases the conformation with lowest energy does not necessarily represents the native protein structure.

To solve the problem of *“ab initio”* folding large supercomputers or distributed computing platforms are used. As an intermediate steps towards predicted protein structures several different approach were proposed like contact map predictions (CMAPpro [42]), usage of discrete grids (TOUCHSTONE [43]) or combining the proteins from available small 3D predicted protein fragments (ROSETTA [44]). The constant increasing power of standard desktop computers and more sophisticated algorithms allow currently (in 2013) to perform in reasonable time sampling for large peptides of size up to 40 - 50 amino acids however the time of prediction increases exponentially with the increase of peptide length [45] [46] [47]

5. Improvement and benchmarking of fold recognition methods

5.1. CASP

Critical Assessment of protein Structure Prediction, or CASP, is a worldwide experiment performed by protein predictors for evaluation of structure prediction which take place every two years since 1994 [48]. The primary goal of CASP is to help advance the methods of identifying protein three-dimensional structure from sequence alone. CASP provides research groups and their software with an opportunity to objectively test prediction methodology and delivers an independent assessment of the state of the art in protein structure modeling to the research community and software users.

In order to ensure that no predictor can have information about a protein's structure under evaluation, predictors, organizers and assessors don’t know the structures of the target proteins until the end of experiment. Targets for structure prediction are unreleased or on-hold X-ray crystallography and NMR spectroscopy provided mainly by structural genomics and proteomics centers. The difficulty of the target protein is determined by the sequence based comparison methods to the existing protein structures. If the given sequence is found to be related by common descent to a protein sequence of known structure, comparative protein modeling may be used to predict the tertiary structure. Such targets are classified as easy ones. Templates that can be found using threading or profile methods are classified as fold recognition targets. Third class

is “*ab initio”* targets for which it is difficult to find the template using above methods so “*de novo”* protein structure prediction must be applied[49, 50].

The primary method of evaluation in CASP is a comparison of the predicted model α-carbon positions with those in original structure. The comparison is shown visually by cumulative plots of distances between pairs of equivalents α-carbon in the alignment and the score describing percentage of well-modeled residues in the model called GDT-TS (Global Distance Test - Total Score)[51] is assigned. The GDT score is calculated as the largest set of amino acid residues alpha carbon atoms in the model structure falling within a defined distance cutoff of their position in the experimental structure.

Free modeling (template-free, or de novo) is also evaluated visually by the assessors however in difficult cases with high RMSD the assessment is not straightforward. CASP provides with many other scoring functions. Some of them perform better for local and some for global alignment. Other doesn’t take into account coiled regions. Other methods push more emphasis on correct fold and overall backbone packing then alignment itself. There is no easy answer to question which methods are the best but when taking into account specific prediction categories and target groups one can better compare different methods and check how they are progressing every year.

The overall performance of protein prediction is strongly related to the predictor experience and manuals corrections made. The prediction groups using similar tools may submit slightly different models of the same target. To avoid human factor influence on the production methods the fully automated version of experiment was proposed.

5.2. CAFASP

The aim of CAFASP (Critical Assessment of Fully Automated Structure Prediction) was to assess the performance of methods without the user intervention that several groups used in their CASP submissions. Although currently still human predictors are superior to automated ones[52] it provides an indication of the performance of the methods alone. This information may aid scientists in choosing which programs they wish to use and in evaluating the reliability of the programs when applied to their specific prediction targets.
To avoid submission of manually “curated” models the predictions must have been sent automatically to the evaluation center in very short time span (several minutes).

The experiment runs once every two years in parallel with CASP and recently has been has been incorporated into the CASP. Results for humans and server are evaluated separately in different categories. In contrast to continues benchmarking techniques like

LiveBench[34] and EVA[53], which run weekly against new proteins published in the Protein Data Bank, CAFASP generates much less data.

5.3. LIVEBENCH

LiveBench[34] is a continuously running benchmark project for assessing the quality of protein structure prediction and secondary structure prediction methods. The main advantage of LiveBench is that it is run continuously and unlike the related CASP and CAFASP experiments, LiveBench is intended to study the accuracy of predictions that would be performed by non-expert users of publicly available prediction methods. Live bench is designed to assess the tertiary and secondary structure predictions.

The experiments like CASP and CAFASP can point out best predicting research groups and methods but one of the drawbacks of both experiments is the limited number of protein targets used to assess the quality of groups or prediction methods. LiveBench project follows the CAFASP ideology, but its goal is to overcome the problem of a limited number of targets by selecting a large number of prediction targets through weekly scanning of the protein structure database PDB for novel proteins. Similarly to CAFASP, LiveBench uses sequences of newly released but published structures and assumes that the evaluated servers do not utilize information form those structures for databases incorporated in their algorithms. Immediate availability of the structures and instant predictions allows reducing risk of such events.
LiveBench assesses publicly available fold-recognition servers and it is designed to provide researchers with constant information about quality of prediction and performance of methods used in experiment.

III. Applications profile-profile comparison methods

There are many purposes of profile-profile sequence comparison in molecular biology. They can be used to identification sequences features such as active sites, post translational modification sites, corresponding gene-structures, reading frames, distributions of introns and exons, regulatory elements, domains, secondary structure, metabolic paths and activates, structure prediction and revealing the evolution and genetic diversity of sequences and organisms. The sensitivity of those methods extends those applications for sequences very distant in their evolution for which the fold recognition and sequence comparison could not be performing even 10 years ago using sequence to sequence analysis.

1. Gene identification and detection of distinct homologues

In the publication about detection of chalcone synthase sequences encoded in yellow lupin[1] profile-profile fold recognition methods were used to detect two full copies of cDNA sequences encoding chalcone synthase (CHS) from Lupinus luteus root. Sequence alignment to distant homologues and phylogenetic studies of chalcone synthases from 54 other plant species as well as distant homologue from Deinococcus radiodurans revealed the possibility that lupin chalcone synthase is encoded by multigene family of at least two distinct genes evolved by gene duplication about 16 million years ago. Additionally application of Structure Prediction Metaserver[31] allowed creation of consensus multiple alignment of several profile-profile comparison methods. The alignment of 57 protein sequences was then used to perform of molecular model and detection of 21 amino-acids in chalcone synthase catalytic pocket to confirm that new discovered sequences are functional chalcone synthases. The alignment of profile-profile methods was also used in molecular phylogeny calculations as a guide tree for topology estimation of chalcone synthase tree. As the actual number of possible topologies for 57 sequences was very high, which makes the analysis computationally expensive the tree based on alignment of chalcone synthases detected by profile-profile methods was used to choose the most probable tree topology based on distances between most distant homologues detected and the correct choice was confirmed by a bootstrap test.

2. Detection of domain boundaries and modeling of complex proteins

In the work describing structure and function of tenascin-C gene[2], usage of sensitive profile-profile sequence comparison analysis allowed to detect the order of functional elements in large multidomain tenascin-C protein, all variable part of a molecule as well

as all isoforms of the protein. The number of putative fibronectin repeats was corrected. Also previously not identified HSP33 domain with was described.

In order to identify domain boundaries and the homologs of the Tn-C domains, the Gene Relate Sequence Database (GRDB)[54]. For the corresponding Tenascin-C elements, the characteristic profiles were computed as well as for every protein families collected from Pfam, COG, the PDB7 and from other genomic sources. The comparison of the target families with about 100,000 other families was performed using Meta-BASIC program[25]. Finally, the models of Tenascin-C domains, we searched the PDB database to find distant homologs with known tertiary structure. The sequences of each identified domain were used as a query. The molecular modeling was performed with the MODELER program [55].

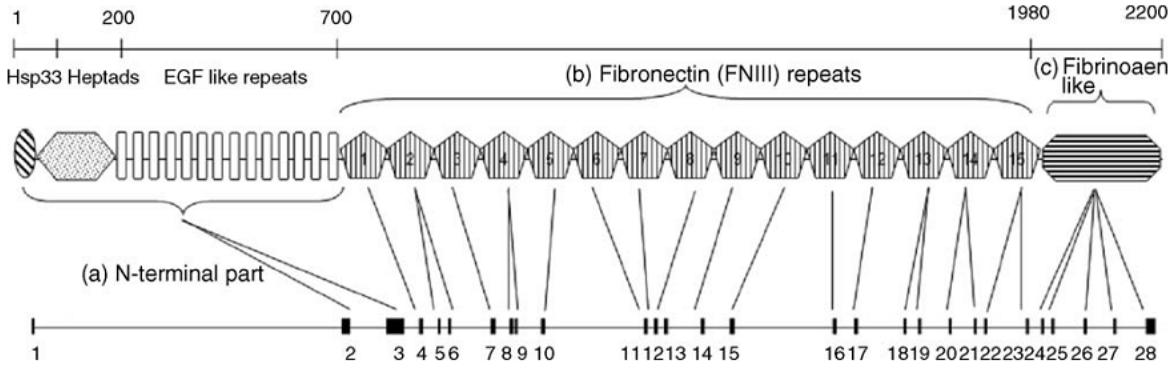

Figure 1. Schematic analysis of Tn-C polypeptide sequence (1-2200 aa). Pictograms represent the repeats involved in the main parts: (a) N-terminal part consisted of HSP33 (diagonal), heptads region (dots) and EGF-like (plain), (b) fibronectin III repeats

To evaluate the local environment and interresidue contacts we used the Verify3D program (Carugo & Pongor, 2002). It provides assessment of the structures on the residue level, which enables the user to locate the parts of a protein that are likely to have the correct con-formation or to look for misfolded regions. The program uses information about the local secondary structure, sol-vent accessibility and the fraction of side-chain area that is covered by polar atoms. We decided to take advantage of this feature and apply Verify3D to improve the incorrectly folded model sections. The alignment in the regions poorly scored by Verify3D was corrected. Oligonucleotide sequences of Tn-C were taken from GEO Profiles database (Barrett et al., 2005) by single blast run with genomic Tn-C sequence.

3. Evolutionary analysis and protein-ligand interaction

In the attempt to characterize evolution and ligand interaction of CRE1 protein[56] the standard methods such as BLAST failed to find close homologs of target gene as well distant structure representative of the CHASE domain. Very often domains may become very diverse during evolution but still possess the same fold which sometimes can be detected using more sensitive tools. Once again the GRDB software[54] was used to identify distant homologue sequence of the CHASE domain and perform model of CRE1 protein.

By applying distant homology detection it was possible to describe CHASE domain as similar to the photoactive yellow protein-like sensor domain. The active site pocket and amino acids that are involved in receptor–ligand interactions was identified. With help of such sensitive tools it was shown that fold evolution of cytokinin receptors is very important for a full understanding of the signal transduction mechanism in plants.

The key future of algorithm of GRDB software is that unlike many other methods used in fold recognition, it does not require any information about any native structure of protein in search. The initial search was performed using truncated CRE1a protein. Using two discovered proteins with known fold which posess sensor kinase activity and which praciptitate in signaling by two-component system the consensus alignment was performed. It confirmed that CHASE, CACHE (CHEmotaxis receptors) and the PAS domains are phylogenetically related.

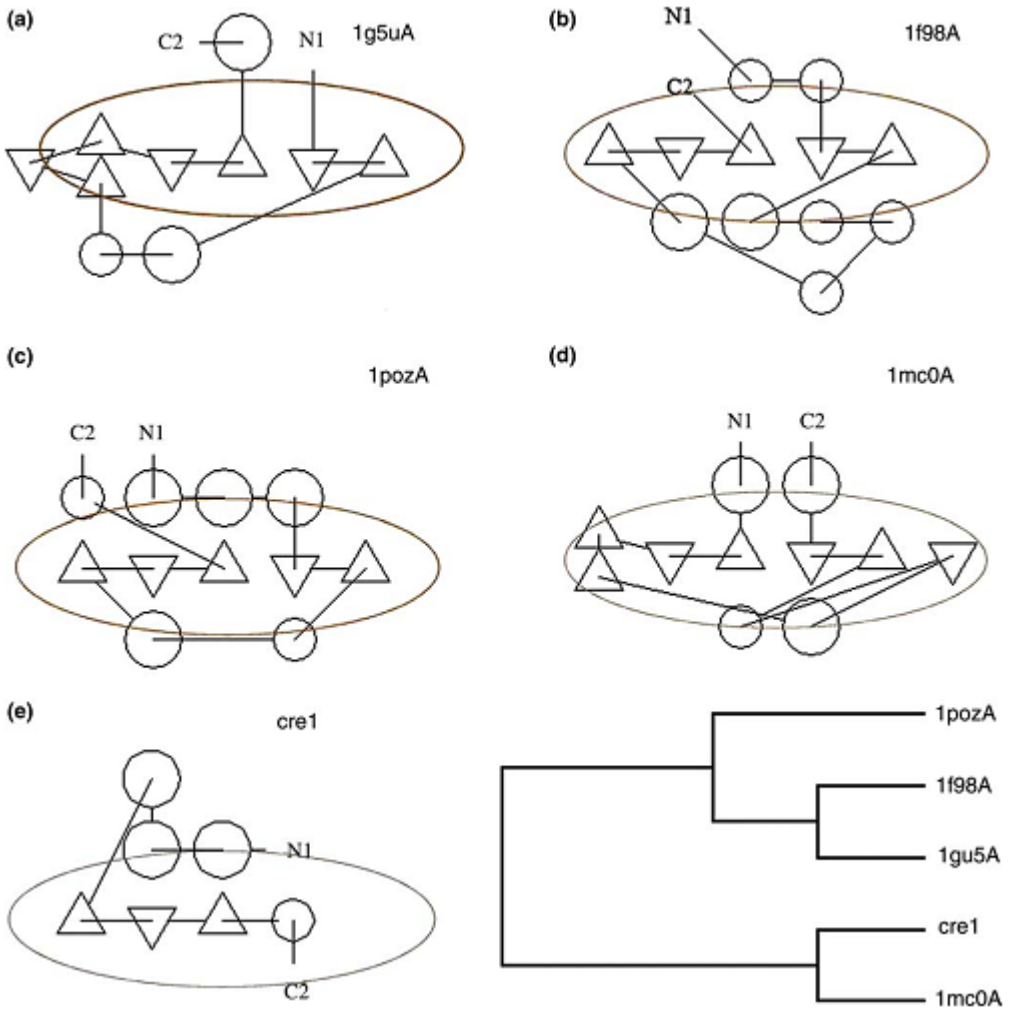

Figure 2. Topological representation and fold evolution of PYP-like family. The cartoon representation of topology of selected related sensing domains of: (a) Profilin; (b) PAS; (c) CACHE; (d) GAP; and (e) CHASE. The triangular symbols represent beta strand, circular symbols represent helices. The direction information for strands is indicated by up and down pointing triangles. Common structural motif in all structures is circled red. Cartoons were performed using the TOPS program. Dendrogram is calculated using PRIDE based on the measure of the degree of similarity between proteins and shows distances between structurally diverged 3D structures. The evolutionary distant sensing domain ACT was used as an outgroup.

Multiple sequence alignment between the consensus sequence of the CHASE domain and the two families shows conservation of sequence pat-terns and secondary structure despite the low amino acid identity. Molecular modeling and docking were performed to examine the possibility of ligand binding. Information about known ligands bound to related proteins was used as initial guide.

The docking of model and related structures was performed. The docked ligand was entirely buried. Molecular modeling confirmed the importance of threonine 278 for the catalytic activity of the enzyme.

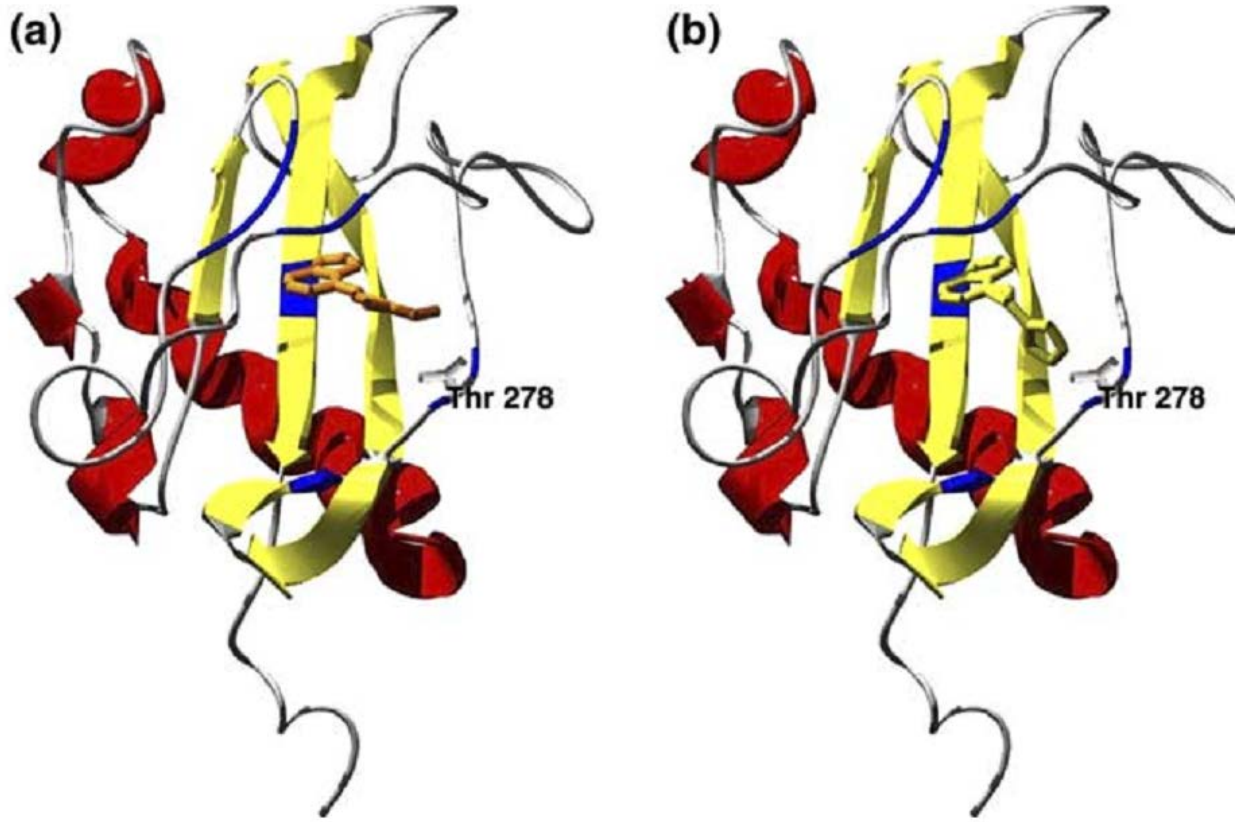

Figure 3. Molecular model of CRE1 receptor from A. thaliana docked with (a) trans-zeatin, (b) kinetin. The yellow and red colors indicate strands and helices, respectively. Ligand is colored in orange. The part of the chain whose residues are at contact distance with ligand is colored in blue. The visible side chain belongs to threonine 278 whose mutation is responsible for loss of function.

IV. Implementations of profile-profile comparison methods

1. The PDB Preview

The Protein Data Bank (PDB) is a central repository where biological macromolecules obtained by X-ray crystallography and NMR spectroscopy are stored and available on the Internet. Not all the entries in the database are publicly available. A new structure can be deposited as “on-hold” entry, non-accessible for public before final release. To allow scientists early access to the 3D structure, it is possible in most cases to generate relatively accurate automatically created computational models. Creation of a database for those on-hold entries that cannot be obtained with standard homology modeling tools as the newly deposited structures often differs significantly in sequence from those already deposited in PDB. Profile-profile Fold Recognition methods allows for automatic generation of models for such “difficult” cases.

The PDB Preview [57] service every week scans, the newly deposited on-hold entries in the Protein Data Bank [58] are scanned to select those that correspond to proteins with no significant sequence similarity to any protein of known structure and perform Fold Recognition using BioinfoBank metaserver and then generates “consensus” models using using the 3D-Jury method [54].

Each prediction is automatically incorporated into the PDB Preview database. The on-hold sequences are described by three different scoring mechanisms: the BLAST E-Value, the PDB BLAST E-Value and 3D-Jury score.

The models with BLAST E-value to the PDB sequence above 0.001 are considered as homology-modeling targets. The PDB Blast E-value to the PDB database clustered at 70% sequence identity using the 3D – Hit [59] and masked with low complexity filters. Usually targets with score below 0.002 are considered as simple Fold Recognition targets.

The 3D-Jury score is obtained with default parameters. Usually 3D-Jury predictions with scores above 50 correspond to essentially correct predictions, meaning that the overall folds of the predicted models are structurally similar to the corresponding experimental structures.

Additionally, the PDB-Preview highlights those 3D-Jury predictions that are regarded as confident but cannot be obtained with a significant score by PDB-Blast. These targets are potentially most interesting cases. The system periodically removes from the PDB-Preview database those PDB entries whose status has changed from “on-hold” to “available”. At this point, the accuracy of the stored predictions can be evaluated, because the experimental 3D coordinates become available. The second section of the

PDB-Preview database corresponds to predictions of previously on-hold entries that have been released.
The PDB-Preview high-scoring Fold Recognition predictions provide biologists with relatively accurate 3D models for not yet released PDB shortly after they are deposited in the PDB, and well before the experimental structure is released. This allows researchers to perform studies that normally are possible only after official 3D coordinates are available. Also the theoretical model allows crystallographers to verify correctness of the new released structure[60]. Additionally the resulting PDB-CAFASP analysis provides computational biologists with a continuous blind evaluation of their methods thus effectively extending other benchmarking experiments such as LiveBench [61] and CAFASP [62].

2. Gene Relational DataBase (GRDB)

The profile-profile comparison was also implemented in GRDB service[63]. Gene Relational DataBase GRDB is the web bases system which contains the characteristic protein sequence profiles for many protein families classification resources such as PFAM [64], CDD [65] and from COGs [66] and also for representatives of structure profiles genereted from PDB[58] or from other genomic sources. GRDB performs the comparison of the target family with 100,000 other families, using Meta BASIC profile-profile comparison methods. It is possible because profile-profile methods in contrast to many other Fold Recognition methods, does not require any information about native structure to perform the comparison. In contrast to pure Meta Basic method the GRDB conducts additional simple PSI-Blast search procedure to include additional information from amino acid sequences translated from open reading frames of unfinished genomes which increases the level of information in the profile.

GRDB was successfully used for comprehensive classification of nucleotidyltransferase fold proteins and identification of novel families and their representatives in human [67]. The strategy used in typical approach of this purpose is usually to collect sequences of know representatives of given family and then using GRDB for searching of missing members. The researcher should note that most trusted results are obtained with the Meta-BASIC[25] score grater then 40. According to rigorous structural criteria used in LiveBench[34] experiment with this score GRDB finds less then 5% of false positive results. It also should be noted then in results with Meta-BASIC score higher then 20 it is possible to find some sequences that are correct. The next step is to find the links between potential representatives of families and deep analysis o sequence alignments and secondary structure of candidates. The last step of the typical search strategy may be analysis of molecular models built based on GRDB predictions.

Currently the following databases are available: CDD[65], PFAM[64], and PDB90[58] (representatives from PDB filtered at 90% of sequence identity). Meta-BASIC calculates connections between 10 340 PFAM, 4852 KOG[68] and 4872 COG[66] families and 20 540 proteins of known structure. The final HTML output is given right away or is send back to the user by email if the computation takes more time.

V. Conclusions

The practical implementations and applications of protein structure prediction are now more important than ever when the massive amounts of protein sequence data are produced by modern large-scale DNA sequencing projects. Despite the efforts in structural genomics, the output of experimentally determined protein structures form X-ray crystallography and NMR spectroscopy are still expensive and time consuming.

The profile-profile comparison methods are most accurate and successful methods so far, especially when there exists a structure template to the target[43]. They are usually superior to other methods because they can pick up possible homologous structure templates even when the sequence identity is very low and that profile-profile comparison can align the sequence to the structure template more accurately, producing more accurate structure models[43]. As more and more novel sequences are produced from the genome projects, the profile-based methods can be expected to become even more sensitive. Fold assignments that were traditionally accomplished from threading methods can be successfully done with comparative modeling instead.

The statistics of Protein Data Base [69] shows that in recent years, only a limited number of completely new protein folds from several thousand new structures are deposited to the databases. In most cases single-domain proteins of up to 200 residues can be aligned to a protein in the PDB with an average RMSD less than 5Å and an average coverage of 70%. These observations imply that in principle, if a reasonably good template can be identified the good quality prediction can be made using profile-profile methods. On the other hand *"Ab Initio"* based methods can still be expected to play an important role in identifying new folds as the accuracy of these methods increase.

The addition of predicted secondary structure to conventional sequence profiles is able to boost the sensitivity of profile-profile comparison methods substantially. The increase of sensitivity of such improved hybrid threading algorithm should eventually result in an increase of specificity. Currently best way to boost the specificity of predictions is the application of consensus methods. This will increase not only the chance that the comparative modeling can assign the fold correctly but also the likelihood that the fold identified is more structurally similar to the target, thus increasing the accuracy of the structural model.

Despite the fact that there might be a limit to which sequence comparison methods can align sequence to structure when the sequence identity is low, further improvement in the sequence-structure alignment can also improve the accuracy of the structure models. The current sequence comparison methods can only align a fraction of the residuals that can be aligned in structure alignments[70]. Better alignment can significantly improve the accuracy of the structure models.

Another way of improving quality of structural models might be refinement made by molecular dynamics (MD) with accurate all-atom physical potentials. The most severe obstacle of the application in MD in protein structure prediction has been the long time which takes for the protein to fold from the completely unfolded states. In homology based methods the energy barriers encountered in the course of folding are removed. If the simulation starts from the near native, the MD simulation perhaps can reach the native structures much easily. The constant increase of computational power, cloud and distributed computing projects like FoldIt! [71] Can bring new perspective for protein structure prediction as well as for increase of profile-profile methods performance.

VI. Figures:

VII. References:

1. Narozna, D., et al., *Two sequences encoding chalcone synthase in yellow lupin (Lupinus luteus l.) may have evolved by gene duplication.* Cell Mol Biol Lett, 2004. **9**(1): p. 95-105.
2. Pas, J., et al., *Analysis of structure and function of tenascin-C.* Int J Biochem Cell Biol, 2006. **38**(9): p. 1594-602.
3. Sanger, F., *The arrangement of amino acids in proteins.* Adv Protein Chem, 1952. **7**: p. 1-67.
4. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
5. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
6. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX.* Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2 3.
7. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* J Mol Biol, 2000. **302**(1): p. 205-17.
8. Vintsyuk, T.K., *Speech discrimination by dynamic programming.* Kibernetika 1968(e): p. 81–88.
9. Wagner, R.A. and M.J. Fischer, *The String-to-String Correction Problem.* J. ACM, 1974. **21**(1): p. 168-173.
10. Meyer, I.M. and R. Durbin, *Comparative ab initio prediction of gene structures using pair HMMs.* Bioinformatics, 2002. **18**(10): p. 1309-18.
11. Schwartz, R.M. and M.O. Dayhoff, *Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts.* Science, 1978. **199**(4327): p. 395-403.
12. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
13. Sankoff, D., *Matching sequences under deletion-insertion constraints.* Proc Natl Acad Sci U S A, 1972. **69**(1): p. 4-6.
14. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.
15. Gotoh, O., *An improved algorithm for matching biological sequences.* J Mol Biol, 1982. **162**(3): p. 705-8.
16. Altschul, S.F. and B.W. Erickson, *Locally optimal subalignments using nonlinear similarity functions.* Bull Math Biol, 1986. **48**(5-6): p. 633-60.
17. Gribskov, M., A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins.* Proc Natl Acad Sci U S A, 1987. **84**(13): p. 4355-8.
18. Brown, M., et al., *Using Dirichlet mixture priors to derive hidden Markov models for protein families.* Proc Int Conf Intell Syst Mol Biol, 1993. **1**: p. 47-55.
19. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

20. Henikoff, S. and J.G. Henikoff, *Embedding strategies for effective use of information from multiple sequence alignments.* Protein Sci, 1997. **6**(3): p. 698-705.
21. Marchler-Bauer, A., et al., *CDD: a Conserved Domain Database for protein classification.* Nucleic Acids Res, 2005. **33**(Database issue): p. D192-6.
22. Schaffer, A.A., et al., *IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices.* Bioinformatics, 1999. **15**(12): p. 1000-11.
23. Ohlson, T., B. Wallner, and A. Elofsson, *Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods.* Proteins, 2004. **57**(1): p. 188-97.
24. Rychlewski, L., et al., *Comparison of sequence profiles. Strategies for structural predictions using sequence information.* Protein Sci, 2000. **9**(2): p. 232-41.
25. Ginalski, K., et al., *Detecting distant homology with Meta-BASIC.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W576-81.
26. Ginalski, K., et al., *ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure.* Nucleic Acids Res, 2003. **31**(13): p. 3804-7.
27. Rychlewski, L., B. Zhang, and A. Godzik, *Fold and function predictions for Mycoplasma genitalium proteins.* Fold Des, 1998. **3**(4): p. 229-38.
28. Gonnet, G.H., M.A. Cohen, and S.A. Benner, *Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix.* Biochem Biophys Res Commun, 1994. **199**(2): p. 489-96.
29. Jaroszewski, L., et al., *FFAS03: a server for profile--profile sequence alignments.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W284-8.
30. Benson, D.A., et al., *GenBank: update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D23-6.
31. Bujnicki, J.M., et al., *Structure prediction meta server.* Bioinformatics, 2001. **17**(8): p. 750-1.
32. Fischer, D., *Hybrid fold recognition: combining sequence derived properties with evolutionary information.* Pac Symp Biocomput, 2000: p. 119-30.
33. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
34. Bujnicki, J.M., et al., *LiveBench-1: continuous benchmarking of protein structure prediction servers.* Protein Sci, 2001. **10**(2): p. 352-61.
35. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A new approach to protein fold recognition.* Nature, 1992. **358**(6381): p. 86-9.
36. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-70.
37. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.
38. Peng, J. and J. Xu, *RaptorX: exploiting structure information for protein alignment by statistical inference.* Proteins. **79 Suppl 10**: p. 161-71.

39. Wu, S. and Y. Zhang, *MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information.* Proteins, 2008. **72**(2): p. 547-56.
40. Yang, Y., et al., *Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates.* Bioinformatics, 2011. **27**(15): p. 2076-82.
41. Levinthal, C., *How to fold graciously*. 1969, University of Illinois Press, Urbana, IL. p. 22-24.
42. Cheng, J., et al., *SCRATCH: a protein structure and structural feature prediction server.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W72-6.
43. Zhang, Y., A. Kolinski, and J. Skolnick, *TOUCHSTONE II: a new approach to ab initio protein structure prediction*. 2003, The Biophysical Society. p. 1145.
44. Simons, K.T., et al., *Ab initio protein structure prediction of CASP III targets using ROSETTA*. 1999, Wiley Online Library. p. 171-176.
45. Kubelka, J., J. Hofrichter, and W.A. Eaton, *The protein folding "speed limit"*. 2004, Elsevier. p. 76-88.
46. Thirumalai, D., et al., *Theoretical perspectives on protein folding*.
47. Prigozhin, M.B. and M. Gruebele, *Microsecond folding experiments and simulations: a match is made*, Royal Society of Chemistry. p. 3372-3388.
48. Moult, J., et al., *A large scale experiment to assess protein structure prediction methods*. 1995, Wiley Online Library. p. ii-iv.
49. Tress, M.L., I. Ezkurdia, and J.S. Richardson, *Target domain definition and classification in CASP8.* Proteins, 2009. **77 Suppl 9**: p. 10-7.
50. Zhang, Y. and J. Skolnick, *The protein structure prediction problem could be solved using the current PDB library.* Proc Natl Acad Sci U S A, 2005. **102**(4): p. 1029-34.
51. Zemla, A., *LGA: A method for finding 3D similarities in protein structures.* Nucleic Acids Res, 2003. **31**(13): p. 3370-4.
52. Kryshtafovych, A., et al., *Assessment of the assessment: Evaluation of the model quality estimates in CASP10.* Proteins.
53. Rost, B. and V.A. Eyrich, *EVA: large-scale analysis of secondary structure prediction.* Proteins, 2001. **Suppl 5**: p. 192-9.
54. von Grotthuss, M., et al., *Application of 3D-Jury, GRDB, and Verify3D in fold recognition.* Proteins, 2003. **53 Suppl 6**: p. 418-23.
55. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.
56. Pas, J., et al., *Structure prediction, evolution and ligand interaction of CHASE domain.* FEBS Lett, 2004. **576**(3): p. 287-90.
57. Fischer, D., J. Pas, and L. Rychlewski, *The PDB-Preview database: a repository of in-silico models of 'on-hold' PDB entries.* Bioinformatics, 2004. **20**(15): p. 2482-4.
58. Sussman, J.L., et al., *Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules.* Acta Crystallogr D Biol Crystallogr, 1998. **54**(Pt 6 Pt 1): p. 1078-84.

59. Plewczynski, D., et al., *3D-Hit: fast structural comparison of proteins.* Appl Bioinformatics, 2002. **1**(4): p. 223-5.
60. Bujnicki, J.M., et al., *Errors in the D. radiodurans large ribosomal subunit structure detected by protein fold-recognition and structure validation tools.* FEBS Lett, 2002. **525**(1-3): p. 174-5.
61. Rychlewski, L. and D. Fischer, *LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction.* Protein Sci, 2005. **14**(1): p. 240-5.
62. Fischer, D., et al., *CAFASP-1: critical assessment of fully automated structure prediction methods.* Proteins, 1999. **Suppl 3**: p. 209-17.
63. Pas, J., et al., *GRDB – Gene Relational DataBase.* Bioinfobank Library Acta, 2011. **2659**.
64. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res. **38**(Database issue): p. D211-22.
65. Marchler-Bauer, A., et al., *CDD: conserved domains and protein three-dimensional structure.* Nucleic Acids Res. **41**(Database issue): p. D348-52.
66. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Res, 2000. **28**(1): p. 33-6.
67. Kuchta, K., et al., *Comprehensive classification of nucleotidyltransferase fold proteins: identification of novel families and their representatives in human.* Nucleic Acids Res, 2009. **37**(22): p. 7701-14.
68. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**: p. 41.
69. Kihara, D. and J. Skolnick, *The PDB is a covering set of small protein structures.* J Mol Biol, 2003. **334**(4): p. 793-802.
70. Sauder, J.M., J.W. Arthur, and R.L. Dunbrack, Jr., *Large-scale comparison of protein sequence alignment algorithms with structure alignments.* Proteins, 2000. **40**(1): p. 6-22.
71. Khatib, F., et al., *Algorithm discovery by protein folding game players.* Proc Natl Acad Sci U S A. **108**(47): p. 18949-53.

VIII. Published articles: