國立交通大學

生物資訊研究所

碩士論文

藉由序列與結構比對辨認關係較遠的同源序列

RDHSP – Recognize Distant Homologues by Sequence-structure profile alignment

研究生：王志淵

指導教授：黃鎮剛　教授

中華民國九十三年七月二十日

DHES － 藉由序列與結構比對辨認關係較遠的同源序列

Recognize Distant Homologues by sequence-structure profile alignment

研究生：王志淵　　Student：Chic-Yuan Wang

指導教授：黃鎮剛　　Advisor：Chih-Ming Hwang

國立交通大學

生物資訊研究所

碩士論文

A Thesis
Submitted to Department of Bioinformatics
College of Biological Science and Technology
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in
Bioinformatics
July 2004

Hsinchu, Taiwan, Republic of China

中華民國九十三年七月

學生：王志淵　　　　　　　　　　　　　　　指導教授：黃鎮剛教授

國立交通大學生物資訊研究所 碩士班

摘　　　要

RDHSP 為一蛋白質摺疊辨識法,其目的為給予一個序列然後計算出這個序列最有可能摺疊成為哪一個結構，胺基酸序列和立體結構在空間中的位置做排列，藉著適當的計分方法，計算這樣排序的得分，憑著得分的高低，判斷序列是不是會摺疊成這樣的立體結構，這種計算序列與結構之間的排序過程稱也稱作為 Threading。

RDHSP 利用待測蛋白質在 19 種「環境」(environment)(二級結構,量測蛋白質中被包埋起來的 side chain 面積疏水性及非疏水性,接觸能量,鄰近的胺基酸個數)作用的結果來分別描述此一蛋白質不同部位的結構，在此所提到的「環境」一詞，係指不同蛋白質 residue 間的接觸形式,接著建立在 20 種胺基酸在不同 environment 中的觀察結果，將所得資料與已知蛋白質結構或序列的資料庫比對取一胺基酸序列對上述 3D profile 進行排序(alignment)，比對的計分標準則依據此一序列和 3D profile 中所描述的結構相容性高低來判斷。

RDHSP-Recognize distant Homologues by Sequence-structure Profiles comparison

Student: Chih-Yuan Wang Advisor: Dr. Jenn-Kang Hwang

Institute of Bioinformatics

National Chiao Tung University

Abstract

The RDHSP (**R**ecognize **D**istant **H**omologues by **S**equence-structure **P**rofile comparison) is the protein fold recognition using the threading method. For a given target protein sequence and a template structure, RDHSP guarantees to find a globally optimal threading alignment between the two. RDHSP is based upon the premise that structure is better conserved than sequence. Every residue in a protein tertiary structure exists in a particular environment that can be described by features such as main-chain conformation, solvent accessibility, and contact energy, contact residue numbers. RDHSP employs environment-specific scoring table that offer a more precise and discriminating measure of substitution probabilities. Compare with the popular PSI-BLAST, RDHSP is 4.6%, 22.2% and 21.6% more sensitive in detecting the family, superfamily, fold similarities in the Lindal benchmark.

Key words : fold recognition; threading; globally optimal threading, contact energy, contact residue numbers

A table of contents

Table index

Figure index

Symbol

e_{ij}	The effective contact energy
n_{ij}	The number of residue *i* - residue *j* contacts
n_{io}	The number of residue *i*-solvent contacts
n_{0j}	The number of residue *j*-solvent contacts
n_{00}	The number of solvent-solvent contacts
$Z\ (i\ ,m)$	The environment-specific amino acid statistical score
$P(i)$	The probability of finding an amino acid of type *i* in any environment
$P(i,m)$	The probability of finding an amino acid of type *i* in the environment of type *m*
Q_S	Query sequence
T_S	Template sequence
$E_{structure}$	The scoring function, which is a measure of the match of a query sequence Q_S and target structure T_S
$n(i,m)$	The number of amino acids of type *i* in the environment *m*
$s(i,m)$	The score of amino acids of type *i* in the environment *m*

1. Introduction

Threading is a method for finding a tertiary structure that matches a primary sequence of interest. It begins with a database of known structures aligns the sequence of interest against the structures, and evaluates the alignments for significant fit. It is based upon the supposition that there is a limited set of protein folds, and that a given primary sequence can be grouped into one type of fold or another, even if there is not extensive primary sequence similarity. Each residue in a protein tertiary structure stays in a particular structural environment, which can be described by the structural features such as main-chain conformation, solvent accessibility. It has been demonstrated by structural environments and each environment has a distinct substitution pattern [1]. Thus, environment-specific substitution tables offer a more precise discriminating measure of substitution probabilities; compared to traditional substitution tables [2] [3] [4] [5] that do not use any structural information. Environment-specific substitution tables have been used to improve secondary structure prediction [6] [7] and fold recognition.

The dynamic programming method has been adopted by most sequence-structure comparison programs. There are three popular variation of the method: the global alignment algorithm, the local alignment algorithm and the global-local alignment algorithm. A sequence-profile alignment method using a global dynamic programming algorithm was employed to find the minimum of the total score that aligns the query sequence with a template in the template library.

The recognition of homology between protein sequences and known structures provides invaluable information towards understanding the biological behavior and biochemical function of uncharacterized sequences, and enables prediction of three-dimensional structures through comparative. RDHSP is an application for recognizing distant structural homologues of a target sequence by sequence-structure comparison. It assesses the compatibility between a target sequence and structural profiles of all known protein structure families. RDHSP is found to improve most existing fold-recognition methods in sensitive that detect the family, superfamily, fold similarities in the Lindal benchmark.

2. Material and Methods

Overview

Our sequence-structure homology recognition algorithm consists of three stages, as outlined in Figure 1. The first step (broken arrows) is to construct environment-specific amino acid substitution tables by using homologous structure alignments. The second step (continuous arrows) is to generate a database of structural profiles, or a profile library, from individual structures using the environment-specific amino acid substitution tables and gap penalties. The last step (dash-dot arrows) is to align the probe sequence or sequence alignment against each profile in the structure profile library. For each comparison, the statistical significance is evaluated to aid the assessment of sequence-structure compatibility and potential evolutionary relationships.

2.1 About Training Set

Our studies used a training set of 387 proteins (see Table1, which is published as supporting information on the PNAS, Iksoo Chang al. (2001)) from the PDBselect [8] [9] consisting of sequences varying in length from 44 to 1017, with low sequence homology and covering many different three-dimensional-folds according to the Structure Classification of Protein (SCOP) classification [10]. Additional criteria used in selecting the proteins in the training set were follows:

(1)The protein structure was obtained through x-ray crystallography.

(2)The structures were monomeric.

(3)The determined structures missed no more than two amino acid

2.2 Definition of structural environments & The Residue Environment code

Each residue in a protein tertiary structure stays in a particular structural environment, which can be described by the structural features such as main-chain conformation, solvent accessibility. Four groups of structural features have been shown to be useful for describing the local environments of a known structure and improving structural alignments.

(1) Secondary structure:

Four classes were defined: α-helix, β-strand, 3_{10} helix and irregular (coil) structure.

(2) Solvent accessibility:

Two classes were defined: residues with side chain relative accessibilities greater than 7% were defined as accessible, otherwise inaccessible.

(3) Contact energy:

7 patterns and none contact energy.

(4) Contact residue numbers:

6 patterns and none contact residue numbers

The combination of all four features gives 344 [(4*2*7*6)+(4*2*1)] local structural environment in total.

The Residue Environment code :

A: α-helix

B: β-strand

C: 3_{10} helix

D: Irregular structure

E: Accessible (exposed to solvent)

F: Inaccessible (hydrophobic environment)

G: None contact energy & contact number

Cn_1~ Cn_6: Contact residue numbers

Ce_1~ Ce_7: Contact energy

2.2.1 Extracting residue secondary structure from known protein structures

Secondary structure assignments are calculated with the program SSTRUC [D.K.smith, now also part of the PROCHECK suite of programs (Laskowski et al., 1993)]. Sstruc is a program to read in a Brookhaven format file and to calculate torsions and secondary structure assignments. It was originally written by David Smith as a replaceme- nt for the DSSP program of Kabsch and Sander. This program calculates the secondary structural state according to the definition of Kabsch and Sander (1983). It also provides information about the main chain dihedral angles Φ, Ψ and ω. In a color PostScript and an HTML output of JOY [18], repeating elements of secondary structure(α -helix, 3_{10}, π helix, and β -strands) are shown in different colures (Figure 2).

2.2.2 Extracting residue solvent accessibility from known protein structures

The partitioning of residues between a polar aqueous phase and a generally hydrophobic phase (the core of a globular protein) is established to be a major determinant in the process of protein folding. Residues in the solvent-inaccessible core of a protein are more conserved and are thus more useful for identifying distant evolutionary relationships. The program PSA is used to calculate the relative solvent-accessible surface area of all residues in a protein. The program uses an implementation of the algorithm of Lee and Richards(1971). Residues are defined as inaccessible by comparison to an extended conformation, and by default a 7% relative accessibility cut-off (Figure 3) is applied (Hubbard and Blundell, 1987). The cut-off value can be set as a command line argument of JOY.

2.2.3 Extracting residue contact energy from known protein structures

We applied the Miyazawa–Jernigan procedure (MJ) to derive contact energies. Specifically, amino acid residues were represented by the centroids of their side chains (C_α), and two residues were considered to be in contact if the distance between their centroids fell within R_C = 6.5 Å. The numbers of contacts formed between two residues, *i* and *j*, and between them and the solvent molecules (represented by 0) were related to the contact energy by a hypothetical chemical reaction:

$$(i-0)+(j-0) \rightleftarrows (i-j)+(0-0) \qquad (1)$$

The effective contact energy (e_{ij}) is defined as the negative logarithm of the equilibrium constant of the reaction:

$$e_{ij} = -\ln\left(\frac{n_{ij}\, n_{OO}}{n_{iO}\, n_{jO}}\right) \qquad (2)$$

where n_{ij} is the number of residue *i* - residue *j* contacts, n_{io} is the number of residue *i*-solvent contacts, n_{0j} is the number of residue *j*-solvent contacts, n_{00} is the number of solvent-solvent contacts. (Figure 4)

2.2.4 Extracting residue contact numbers from known protein structures

Amino acid residues were represented by the centroids of their side chains(C_α), and two residues were considered to be in contact if the distance between their centro- ds fell within R_C = 6.5 Å.

2.3 Environment-specific amino acid Scoring matrix

In addition to sequence information we also use structural information that can be included in several different ways. We described each position of a protein as being in one of 19 environments. Other researches have developed similar methods e.g. (Ouzounis et al., 1993; Yi & Lander, 1994). The environments in these methods are characterized by properties such as exposed atomic areas and type of residue-residue contacts.

The principles of all these methods are as follows (Figure 5):

1. Reduction of the three-dimensional structure to a one-dimensional string of residue environments. We defined these environments by measuring the area of the side chain that is buried in the protein, contact energy, contact residue nu- mbers and the local secondary structure.
2. A scoring matrix is generated from the probabilities of finding each of the twenty amino acids in each of the environment classes as observed in a database of known structures and related sequences.

2.3.1 Selection of structural alignment

HOMSTRAD (**Hom**ologous **Str**ucture **A**lignment **D**atabase) is a database of protein structure alignments for homologous families. It’s a high-quality database that contains a set of 3D protein structures arranged into families. A subset of the Homstrad database, which consisted of structural alignments of high-resolution structures, was constructed as follows (1) Families of membrane proteins were removed. (2) Only those than 2.5Å were accepted. (3) The highest resolution representatives were selected to ensure that each structure has sequence identity less than 80% to any other structures in the sa- me family.(4)After the first three steps, the families with at least two structures left w- ere retained for substitution calculation. The database HOMSTRAD presently contains 130 protein families and 590 aligned structures. For each family, the database provides a structure-based alignment derived using COMPARER and annotated with JOY in a special format that represents the local structural environment of each amino acid residue. The database is freely available on the World Wide Web at: http://www-cryst.bioc.cam.ac.uk/data/align/.

2.3.2 Calculation of environment-specific amino acid Scoring matrix

The statistical score Z (*i, m*) associated with amino acid *i* in an environment *m* is readily deduced by using the expression:

$$Z\ (i, m)=-\ln [P\ (i, m)/P\ (i)] \tag{3}$$

where *P* (*i*) is the probability of finding an amino acid of type *i* in any environment, and *P*(*i*,*m*) is the probability of finding an amino acid of type *i* in the environment of type *m*. *P* (*i*) and *P* (*i, m*) are determined from knowledge of the sequences and native state structures in our training set. These probabilities were determined from the 387 training set of homologous structure alignment. For each position in the aligned set of sequences, we determined the environment category of the position from the known structure and counted the number of each residue type found at the position with the set of aligned sequenc- es. A residue type was counted only once per position. For example, if there were ten aspartates and one glycine found at a position in a set of aligned sequences, then both the Asp and Gly counters were both incremented by only one. If the number of residues *i* in an environment *m* was found to be zero, the number was increased to one so that *P*(*i*,*m*) was never zero(Figure 6).The environment-specific amino acid scoring matri- x were shown in (Table 2), (Table 3), (Table 4), and (Table 5)

2.4 The Fold Recognition Scoring function

The scoring function is:

$$E_{total} = \omega_{structure} E_{structure} + E_{gap}, \quad \omega_{structure} : \text{weight factor} \tag{4}$$

$$E_{structure}(Q_s, T_s) = \sum_{i} \sum_{m} n(i, m) s(i, m) \tag{5}$$

where $E_{structure}$ is the scoring function, which is a measure of the match of a sequence Q_S and target structure T_S, n(i,m) is the number of amino acids of type i in the environment m, and s(i,m) is the score associated with it.

2.5 Determination of initial gap, extension gap penalties and the weight factor of $\omega_{\text{structure}}$

All weight factors and structure-dependent gap parameters were obtained by optimizing the performance of the method in Fischer’s dataset [31]. The initial gap penalty (ω_0) and the extension gap penalty (ω_1) of the RDHSP method were first optimized by using Fischer’s dataset without the knowledge-based, structure-derived score. The dataset contains 68 probe sequences and 301 library structures. A match occurs when the expected match was ranks as the number one or only below its superfamily members based on SCOP 1.61 classification (i.e., no incorrect folds have a better score than the expected match). The highest success rate is 51/68 when ω_0=1.2 , ω_1=0.7.and $\omega_{\text{structure}}$ =1.3

2.6 Convert known structure into a 1D-3D profile

Generation of a position-dependent comparison matrix known as the 3D profile, ie. defining the probability to find a certain amino acid in a certain position of a given protein. (Figure 8)

2.7 Query sequence – Structure profile alignment

Alignment of a sequence with the 3D profile. The resulting alignment score is a measure of the compatibility of the sequence with the structure described by the 3D profile (Figure 9). A sequence profile alignment method using a global dynamic programming algorithm was employed to find the minimum of the total score that aligns the query sequence with a template in the template library.

3.Result

3.1 Test of threading model

The score of the gapless is given by

$$E_{structure}(Q_s, T_s) = \sum_{i} \sum_{m} n(i, m)s(i, m) \qquad (8)$$

where $E_{structure}$ is the scoring function, which is a measure of the match of a sequence Q_S and target structure $T_{S,}$ n(*i*,*m*) is the number of amino acids of type *i* in the environment *m*, and s(*i*,*m*) is the score associated with it. We take the three largest decoy sets from t- he Prostar website (http://prostar.umbi.umd.edu). For each structure, we sum the gaple- ss threading scores of every residue. If the summed gapless threading score of the nativ- e structure is less than that of the decoy, we consider that our threading model perform- ed correctly in the discrimination of this native-decoy pair. Of the 109 structure-decoy pairs in three different sets, RDHSP correctly detected 94 pairs (Table 2). Further, it is better than residue contact potentials RKBP (81/109; Lu and Skolnick, 2001) and CDF (75/109; Samudrala and Moult, 1998). For the decoy set ig_structal_ hires from Decoy- s'R'us, Figure 8 shows the correlation between the RMSD (root- mean-square-distance) and the score of gapless threading for the decoy set of protein with PDB code 1hda-B. The closer the decoy structure is to the native structure, its s- core is often lower. For th- is protein, the native structure has the lowest score.

3.2 Lindahl Benchmark for Fold Recognition Sensitivity

The Lindahl set was designed to assess the fold recognition sensitivity. It has 976 proteins. Each protein is aligned with the rest 975 ,proteins. There are 555 pairs of proteins in the same family, 434 pairs of proteins in the same superfamily, 321 pairs of proteins in the same fold. The fold-recognition method is tested by checking whether or n- ot the method can recognize the member of same family, or fold as the first rank or wit- hin the top five ranks. The results of RDHSP are compared with several well establish ed methods in Table 3. RDHSP compared to the popular PSI-BLAST, RDHSP is 4.6%, 22.2% and 21.6% more sensitive in recognizing the member of same family, superfamily, and fold.

Results are summarized in Table 7. It can be seen that the performances of RDHSP on all similarity levels are significantly at superfamily/fold levels to any other method. On the family level, all methods except THREADER perform well, and the performance of RDHSP is significantly better than that of the PSI-BLAST. On the superfamily an- d fold levels, RDHSP performs more than 10% better than any other method except FUGUE. It is noticeable that although the focus of RDHSP is on detecting superfamily/fold level similarity, it also performs relatively well in detecting family level similarity, better than PSI-BLAST. RDHSP strength is that unlike other methods, it performs best at superfamily/fold levels. For example, FUGUE performs best at family/superfamily levels and THREADER performs best at fold level.

3.3 A specificity-sensitivity curve at family, superfamily and fold level

We have used two different criteria to analyze the performance of a particular method on Lindahl Benchmark. First we simply examined the fraction of true hits in first and top five ranks, respectively. This is a very intuitive measure, but it tells nothing abo- ut the reliability of the match, i.e a match could be the top rank but still have a very low score as long as all other hits have even lower scores. To overcome this limitation we have used spec-sens plots (Rice & Eisenberg, 1997; Arvestad et al., 1999; Hargbo & Elosson, 1999) as a complementary measure, describing the fraction possible correct hits found as a function of the fraction found hits being correct. The main advantage to this is that it measures the ability of a method to reliably find all pairwise matches in the database. The fraction possible correct hits found, sensitivity, is defined as:

$$\text{Sensitivity(score)} = \frac{\text{TP(score)}}{\text{TP(score)+FN(score)}}$$

where TP(score) is the number of correct hits having a score above score, and FN(score) being the number of correct hits with a score less than score. The specificity measures the probability that a pair of sequences with a score greater than a certain threshold reall- y is a true hit, defined as:

$$\text{Specificity(score)} = \frac{\text{TP(score)}}{\text{TP(score)+FP(score)}}$$

where FP(score) is the number false hits that have a score above score and TP is defined as above. The sensitivity is plotted as a function of specificity, each point corresponding to a certain score. This measure is similar but not identical to the plots described by Par- k et al.(1997) and (1998) where sensitivity, referred to as “fraction of homologous pairs detected”, was plotted against “rate of false positives”.

The specificity-sensitivity curves are shown in Figure 10, Figure 11 and Figure 12 RDHSP performs significantly better than all the other methods at superfamily/fold levels.At the family level, RDHSP achieves the sensitivity of 63.5% and 34.4% at 50% and 99% specificities, respectively, which are roughly 7% better than those of PSI-BLAST. At the superfamily level, RDHSP achieves 25.0% and 11.4% sensitivities at 50% and 99% specificities. RDHSP has much better sensitivity than any other method at superfamily/fold level.

Conclusions

There are some features in RDHSP, (1) using residue contact energy and residue contact numbers to improve environment-specific substitution tables, (2) combined information from both multiple structures and multiple sequences, resulted in the improvement of both homology recognition performance and alignment quality, (3) using sequence-structure profile alignment.

Each residue in a protein tertiary structure stays in a particular structural environment, which can be described by the structural features such as main-chain conformation, solvent accessibility. RDHSP using four groups of structural features (main-chain conformation, solvent accessibility, and contact energy, contact residue numbers) to improve structural alignments.

To further analyze the relative contributions of different term in the structure score, we test our potential in all other five combinations (secondary structure (S) only, secondary structure + solvent accessibility area (ASA), S +ASA+contact energy (Ce), S+ASA+ contact residue numbers (Cn), and S+ASA+ Ce + Cn) without any gap penalty (gapless threading). The results on ProStar benchmark are shown in Table 8. The secondary structure term does not seem to help for alignment accuracy. However, when it was combines with ASA term, the two-term threading score (S+ASA) becomes the better (50.5%). The ASA term appears to be the most important. When (S+ASA) was combine with co- ntact energy or Contact residue numbers, The three- term threading score (S+R +Cn or S+R+Ce) becomes better than two-term (56.9%, 67.0%).