Amino Acids (2010) 39:887–898
DOI 10.1007/s00726-010-0547-x

ORIGINAL ARTICLE

Use of the Burrows–Wheeler similarity distribution to the comparison of the proteins

Lianping Yang · Guisong Chang · Xiangde Zhang · Tianming Wang

Received: 29 August 2009 / Accepted: 25 February 2010 / Published online: 19 March 2010

Abstract In this paper, we present an approach based on Burrows–Wheeler transform to compare the protein sequences. The strings representing amino acid sequences do not reflect the chemical physical properties better, and it is very hard to extract any key features by reading these long character strings directly. The use of the Burrows–Wheeler similarity distribution needs a suitable representation which can reflect some interesting properties of the proteins. For the comparison of the primary protein sequences we convert the protein sequences into digital codes by the Ponnuswamy hydrophobicity index, and for the comparison of the structure of the proteins we adjust the topology of protein structure strings, which are simple but useful representation of the secondary structure of proteins to match the Burrows–Wheeler similarity distribution. At last, some experiments show that the approach proposed in this paper is a powerful and useful tool for the comparison of proteins.

Keywords Burrows–Wheeler transform · Burrows–Wheeler similarity distribution · Phylogenetic tree · Similarity · Hydrophobicity · TOPS string

L. Yang · G. Chang · T. Wang
School of Mathematical Science,
Dalian University of Technology, Dalian 116024, China

L. Yang · G. Chang · X. Zhang (✉)
College of Sciences, Northeastern University,
Shenyang 110004, China
e-mail: zhangxdmath@yahoo.com.cn

L. Yang
e-mail: yangmath@yahoo.cn

Introduction

Proteins are chains of 20 amino acids that are the source of most biological activity and structure in cells. Analysis of amino acid sequences can provide useful insights into the tertiary structures of proteins and their biological functions. The similarity of the two protein sequences may mean the similar function or the similar structure of the two sequences. Many studies have indicated that sequence-based prediction approaches, such as protein interaction based on sequence similarity (Cheng et al. 2009), protein 3D structure prediction based on sequence alignment (Chou 2004), protein subcellular location prediction (Chou and Shen 2008), estimations of amino acid background distribution (Dou et al. 2009, 2010), can provide very timely and useful information and insights into both basic research and drug design and are widely welcomed by the science community. Hence, one of the most important problems is how to measure the similarity between two biological sequences. The first widely accepted solution to this problem is based on alignments. However, sequence alignments seem inadequate for measuring mutations that involve longer segments of sequences because they consider local mutations only. There are many integral properties lost if we use the alignment methods only.

Many alignment-free methods have been introduced recently (Ferragina et al. 2007; Vinga and Almeida 2003; Mantaci et al. 2008; Nandy et al. 2009; Jia et al. 2009; Liu et al. 2007; Feng and Wang 2008b). Graphical representations of proteins have emerged as one kind of alignment-free methods of sequences analysis (Randić et al. 2006; Randić 2007; Liu and Wang 2007; Zhang and Zhang 2000; Liao et al. 2006; Cao et al. 2008). Such representations can make some special patterns visualization. Besides those, there are many alignment-free

methods based on the idea that the more similar two sequences are, the greater the number of the factors shared by two sequences is. One of the oldest approaches based on this idea is the frequencies statistics of the length k words shared by the compared sequences (Blaisdell 1986). In fact, it is proved that the dissimilarity values observed by using distance measures based on words frequencies are directly related to the ones requiring sequence alignment (Blaisdell 1989). So it is often used as a filtration before alignment. Since using the $k-$ words only cannot construct phylogeny tree, the Markov model is often employed to get the information of the biological sequences (Pham and Zuegg 2004; Pham 2007; Kantorovitz et al. 2007; Helden 2004; Dai et al. 2008).

Recently the idea based on text compression technique has been noticed by many researchers (Ferragina et al. 2007). The core idea is that two sequences are considered close if one sequence can be compressed significantly under the condition that the other sequence information is known. A theoretical precursor is a similarity metric based on a notion from the Kolmogorov complexity theory (Li et al. 2004; Li and Vitányi 1997). Since the Kolmogorov complexity is uncomputable, some compressors are introduced to approximate it (Li et al. 2001, 2004; Chen et al. 2004; Cilibrasi et al. 2004).

There are also some other important methods such as Lemple-Ziv (LZ) complexity, Burrows–Wheeler (BW) transform (Mantaci et al. 2007, 2008; Otu and Sayood 2003; Zhang and Wang 2010) which are based on compression algorithm, but do not actually apply the compression. The Burrows–Wheeler Transform (BWT) was introduced by Burrows and Wheeler (1994), and it is a useful tool for textual lossless data compression. The BWT algorithm can transform a string, defined over an alphabet A, to another string which can be compressed easier. The compression is lossless due to the reversible BWT algorithm. To compare the similarity of two sequences, Mantaci et al. (2008) introduced an extension of the Burrows–Wheeler Transform (EBWT) and defined a class of dissimilarity measures. However, they have not considered what is the role of the order relation on the alphabet set A. There are 20 elements in the set A when the sequences are proteins. So there are 20! order relations totally. Now there is a question: which order relation is fit for our problem?

So far, almost all such comparisons, alignment or alignment-free methods are based on the strings. The strings representing the biological sequences, especially the amino acid sequences, do not reflect the chemical physical properties better. It is hard to extract any key features by reading these long character strings directly. We need an appropriate form which is suitable for some special purpose to represent the sequences.

In this paper, we use Burrows–Wheeler similarity distribution (BWSD) (Yang et al. 2009) based on Burrows–Wheeler transform to express the similarity between two protein sequences. There are some distance measures spontaneously followed, such as expectation, entropy, and some other divergence measures of the distribution. When the primary protein sequences are considered, the Xiao and Chou (2007) method is used to convert the protein sequence to digital codes to reflect chemical physical properties better. Since the protein sequences are represented by binary numbers, there are only 2 rather than 20! order relations on the alphabet set A. Hence, we can choose a suitable order relation easier. Further, the use of BWSD is to compare the secondary structure of proteins based on the adjusted TOPS strings which adapt to the BWSD. The results show the validity of the approach in the field of structure comparison. At last, we demonstrate the performance of the BWSD in the field of the construction of phylogenetic trees. Hence, the BWSD is a useful and powerful tool for the comparison proteins.

Burrows–Wheeler transform

Let A be an ordered alphabet. If we denote the set of words over A by A^* and $u, v \in A^*$, then we can define the order of the u and v by the lexicographic order. Hence, the set A^* is a total ordered set. We call u, v *conjugate* denoted by $u \sim v$ if $u = xy$ and $v = yx$ for some $x, y \in A^*$. u is called a *primitive* word if $u = x^n$ implies $u = x$ and $n = 1$.

If $u = a_1 a_2 \cdots a_n \in A^*$ is a primitive word, let $S_u = \{v \in A^* | u \sim v\}$. Then the number of the set S_u is n. Since $S_u \subseteqq A^*$ and A^* is a total ordered set, the set S_u can be sorted. If the input of the BWT algorithm is the word u, then the output of the BWT [denoted by bwt(u)] is the pair (L, I) where L is the sequence of the last character of each word in the sorted set S_u and I is the position of the original word u in the sorted set S_u. That is bwt(u) = (L, I). Denote BWT(u) = L.

For example, let $u = babrbca$, then the sorted set S_u is:

ababrbc
abrbcab
babrbca
bcababr
brbcaba
cababrb
rbcabab

bwt(u) = (*cbarabb*, 3). BWT(u) = *cbarabb*.

There are some conclusions about the BWT. The proofs can be found in Mantaci et al. (2007) and Mantaci et al. (2003).

Proposition 1 *$u \sim v$ if and only if* BWT(u) = BWT(v).

Proposition 2 *$u \sim v^d$ if and only if* $\mathrm{BWT}(v) = b_1 b_2 \cdots b_m$ *and* $\mathrm{BWT}(u) = b_1^d b_2^d \cdots b_m^d$.

Let $C_1 = \{L \in A^* | L = \mathrm{BWT}(u)$, for some $u \in A^*$ and u is primitive$\}$.

Proposition 3 *For any $L = l_1 l_2 \cdots l_n \in C_1$ and integer $I \leq n$, there exists unique $u \in A^*$ such that* $\mathrm{bwt}(u) = (L, I)$.

In fact, the set S_u can be obtained by L and u can be obtained by I from the set S_u. The algorithm about the bwt is represented in Mantaci et al. (2003, 2007, 2008).

Now consider two words $u, v \in A^*$ which are primitive. According to the BWT, we can obtain S_u and S_v. Let L be the sequence of the last character of each word in the sorted multi-set $S_u \cup S_v$ (If $u^* \in S_u$, $v^* \in S_v$ and $u^* = v^*$, then $u^* < v^*$). Label the characters of L which are from the word $u^* \in S_u$ by 0s and label others by 1s. α is the vector consisting of the labels. β is the vector consisting of the positions of the u, v in the set $S_u \cup S_v$. Denote $\mathrm{bwt}(u, v) = (L, \beta)$ and $\mathrm{BWT}(u, v) = L$, $\mathrm{char}(u, v) = \alpha$. Since $v_1 < v_2$ in S_v implies $v_1 < v_2$ in $S_u \cup S_v$, S_v is shuffled in the S_u.

For example, $u = babrbca$, $v = abrcabb$, then the sorted multi-set $S_u \cup S_v$ is: {ababrbc, abbabrc, abrbcab, abrcabb, babrbca, babrcab, bbabrca, bcababr, brbcaba, brcabba, cababrb, cabbabr, rbcabab, rcabbab}.

$$\mathrm{bwt}(u, v) = (ccbbabaraabrbb, (5, 4)),$$
$$\alpha = (0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1).$$

The similarity analysis based on Burrows–Wheeler similarity distribution

In Mantaci et al. (2008) a class of distance measures between words was defined by the colored EBWT transformation. All the distance measures in the family are based on the intuitive idea that the greater the number of factors shared by two sequences u and v is, the smaller the distance between u and v is. To get these distance measures, the $\mathrm{char}(u, v)$ should be segmented at first. Let $P = \{x_1, x_2, \ldots, x_k\}$ be the segment. Then $D_P(u, v) = \sum_{x \in P} |n_x(u) - n_x(v)|$ where the $n_x(u)$ is the number of the 0s in the block x, $n_x(v)$ is the number of the 1s in the block x. An important property of these measures is that it includes, as particular case, the $k-$ word counts Euclidean distance which corresponds to a special choice of the segment method P.

The order relation between primitive words in EBWT is different from the order relation mentioned in BWT. We notice that the new order relation used in EBWT is not very important to compare two sequences while what makes a great difference is the idea that in the sorted list consisting of all the conjugates of the two words, the greater the number of factors shared by two words is, the greater the conjugates mix. As a result we use the original order relation rather than the order relation used in EBWT.

In this paper, we consider the $\mathrm{char}(u, v)$ from another viewpoint. Given two words u, v, consider $\alpha = \mathrm{char}(u, v)$. First, we rewrite α to the form $0^{k_1} 1^{k_2} 0^{k_3} 1^{k_4} \cdots 0^{k_m} 1^{k_{m+1}}$ where i^{k_j} means i repeats k_j times. Note that if $j \neq 1$ and $j \neq m + 1$, then $k_j > 0$. Let t_n be the number of k_j such that $k_j = n$. Let $s = t_1 + t_2 + \cdots + t_n + \cdots$. Note that there are only finite numbers of t_n such that $t_n \neq 0$. The BWSD of u, v is $P\{S_{uv} = k\} = t_k/s, k = 1, 2, 3, \ldots$

For example, $u = babrbca$, $v = abrcabb$, $P\{S_{uv} = 1\} = 5/6$, $P\{S_{uv} = 2\} = 1/6$, $P\{S_{uv} = k\} = 0$ if $k > 2$.

One of the basic ideas of comparing two sequences is that the more similar two sequences are, the greater the number of the factors shared by the two sequences is. Mantaci et al. (2008) pointed out that if the same factor s appears both in u and v, then the conjugates of u and v starting with s are likely to be close in the sorted set $S_u \cup S_v$. That means the greater the scale of mixing of conjugates of u and v in the sorted set $S_u \cup S_v$ is, the more similar u and v are. There are many methods to compare the similarity of the two sequences through considering the scale of the mixing. In this paper, we compute the mixing scale of the two sequences by the BWSD.

Definition 1 The distance of two sequence u, v is $D_M(u, v) = E(S_{uv}) - 1$, where $E(S_{uv})$ is the expectation of the BWSD of u, v.

Definition 2 The distance of two sequence u, v is $D_E(u, v) = -\sum_{k \geq 1, t\_k \neq 0} t_k/s \log_2 t_k/s$

Note that it is the Shannon entropy of the BWSD.

If $u \sim v$, then the BWSD is $P\{S_{uv} = 1\} = 1$, $P\{S_{uv} = k\} = 0$ if $k > 1$. Hence we obtain:

Proposition 4 *If $u \sim v$, then $D_M(u, v) = 0$ and $D_E(u, v) = 0$. In particular, $D_M(u, u) = 0$ and $D_E(u, u) = 0$.*

Since S_{uv}, S_{vu} have the same distribution, the next proposition is obtained:

Proposition 5 *$D_M(u, v) = D_M(v, u)$, $D_E(u, v) = D_E(v, u)$ for any two sequences u, v.*

There are some other measures which can be defined based on the BWSD, such as the relative entropy between the BWSDs of the given sequences and the BWSDs of the random sequences. There are still some other divergence measures in Lin (1991) and Zhang et al. (2009).

Up to now, we have not considered what is the role of the order relation on the alphabet set A. We cannot expect the BWSD to be fixed when the order relation is changed.

Table 1 Three different types for coding amino acids and Ponnuswamy hydrophobicity index

Character	Binary	Decimal	Hydrophobicity index
K	00110	6	5.72
N	01000	8	6.17
D	01001	9	6.18
E	01010	10	6.38
P	01011	11	6.64
Q	01100	12	6.67
R	01101	13	6.81
S	01110	14	6.93
T	01111	15	7.08
G	10000	16	7.31
A	10001	17	7.62
H	10010	18	7.85
W	10100	20	8.41
Y	10101	21	8.53
F	10111	23	8.99
L	11000	24	9.37
M	11010	26	9.83
I	11011	27	9.99
V	11100	28	10.38
C	11110	30	10.93

There are 20 elements in the set A when we consider the protein sequences. So there are 20! order relations totally. Which order relation is suitable for our problem? We solve this problem by representing the amino acids as digital codes which reflect the chemical physical properties of protein better.

The representations of proteins

An appropriate representation of protein sequence can provide important insights into the tertiary structures, functions or some other interesting properties. However, the biological sequences are stored in the computer database system in the form of long character strings in general. It would act like a snail's pace for human beings to read these sequences with the naked eyes. Also, it is very hard to extract any key features by directly reading these long character strings (Xiao and Chou 2007). We do not think that there is any representation of proteins which can reflect all the properties. But we can obtain an appropriate representation to reflect some particular properties we are interested in.

Digital coding methods and some topology structure representation methods are employed to analyze the protein sequences. There are many kinds of models on amino acid digital encoding. Through converting the DNA sequences into digital signals and using a base for representation of the nucleotides, Cristea (2001) converts the codons into the numbers in the range 0–63 and the amino acids into the numbers in the range 0–20. According to his model, the 20 amino acids and the terminator are coded as: F = 0, L = 1, S = 2, Y = 3, end = 4, C = 5, W = 6, P = 7, H = 8, Q = 9, R = 10, I = 11, M = 12, T = 13, N = 14, K = 15, V = 16, A = 17, D = 18, E = 19, G = 20. This model distinguishes each amino acid in the process of encoding amino acid only; however, the chemical physical properties of the amino acid are neglected. To append the chemical physical properties, Feng and Wang (2008a) characterize the amino acids using the acid dissociation constants. According to their model, the 20 amino acids are expressed by vectors, i.e., F = (3/17,5/18), L = (15/17,5/9), S = (11/17,1/3), Y = (10/17,2/9), C = (1/17,1), W = (1/2,16/17), P = (4/17,17/18), H = (2/17,7/18), Q = (7/17,5/18), R = (7/17,1/6), I = (15/17,13/18), M = (12/17, 4/9), T = (1,8/9), N = (5/17,1/18), K = (8/17,1/9), V = (13/17,11/18), A = (14/17,7/9), D = (6/17,5/6), E = (9/17, 2/3), G = (14/17,5/9). To reflect better chemical physical properties and degeneracy, Xiao and Chou (2007) encode the amino acids based on hydrophobic index. The hydrophobic amino acids tend to repel the aqueous environment, and therefore reside predominantly in the interior of proteins. The hydrophobicity is one of the major factors that influence the amino acid substitution during evolution. The coding principle is that the larger the

Table 2 Seven protein sequences and structural specification

Protein	Length	GenBankID	Class	Super family
(HAHU) Human α hemoglobin	142	HAHU	All α proteins	Globin-like
(HAHO) Horse α hemoglobin	142	HAHO	All α proteins	Globin-like
(HBHU) Human β hemoglobin	147	HBHU	All α proteins	Globin-like
(CCPG) Pig cytochrome c	104	CCPG	All α proteins	Cytochrome c
(LEGH) Lupine leghemoglobin	154	P02239	All α proteins	Globin-like
(MYWHP) Sperm whale myoglobin	153	MYWHP	All α proteins	Globin-like
(FGFH) Basic human Fibroblast growth factors	288	NP_001997	All β proteins	Cytokine

Ponnuswamy hydrophobicity index of an amino acid is, the greater its digital code is. The digital codes of the 20 amino acids are listed in Table 1. In this paper, we use Xiao and Chou (2007) method to convert the protein sequence to digital code. Since the digital code reflect the hydrophobicity properties well, we can consider the similarity of two protein sequences from the view of the hydrophobicity.

Table 3 Similarity by wavelet-based method (Trad et al. 2002)

	HAHU	HAHO	HBHU	CCPG	LEGH	MYWHP	FGFH
HAHU	5S	5S	1S1W3N	1W4N	2W3N	2W3N	5N
HAHO	5S	5S	1S1W3N	1W4N	2W3N	2W3N	5N
HBHU	1S1W3N	1S1W3N	5S	1W4N	2W3N	2W3N	5N
CCPG	1W4N	1W4N	1W4N	5S	1W4N	5N	5N
LEGH	2W3N	2W3N	2W3N	1W4N	5S	1W3N	5N
MYWHP	2W3N	2W3N	2W3N	5N	1W3N	5S	1W4N
FGFH	5N	5N	5N	5N	5N	1W4N	5S

Table 4 Similarity matrix by spectral-distortion (Pham 2007)

	HAHU	HAHO	HBHU	CCPG	LEGH	MYWHP	FGFH
HAHU	0	0.0029	0.0006	0.0582	0.0148	0.0023	0.0512
HAHO	0.0029	0	0.0043	0.0750	0.0169	0.0042	0.0546
HBHU	0.0006	0.0043	0	0.0488	0.0133	0.0033	0.0479
CCPG	0.0582	0.0750	0.0488	0	0.0365	0.0562	0.0384
LEGH	0.0148	0.0169	0.0133	0.0365	0	0.0072	0.0112
MYWHP	0.0023	0.0042	0.0033	0.0562	0.0072	0	0.0350
FGFH	0.0512	0.0546	0.0479	0.0384	0.0112	0.0350	0

Table 5 Similarity matrix by D_M

	HAHU	HAHO	HBHU	CCPG	LEGH	MYWHP	FGFH
HAHU	0	0.3181	0.4393	1.8080	0.5346	0.5387	0.6220
HAHO	0.3181	0	0.4899	1.8488	0.5836	0.5393	0.6262
HBHU	0.4393	0.4899	0	1.8866	0.5347	0.4958	0.6285
CCPG	1.8080	1.8488	1.8866	0	1.6909	1.9283	2.0467
LEGH	0.5346	0.5836	0.5347	1.6909	0	0.5233	0.6842
MYWHP	0.5387	0.5393	0.4958	1.9283	0.5233	0	0.6905
FGFH	0.6220	0.6262	0.6285	2.0467	0.6842	0.6905	0

Table 6 Similarity matrix by D_E

	HAHU	HAHO	HBHU	CCPG	LEGH	MYWHP	FGFH
HAHU	0	1.0443	1.2751	2.4541	1.4264	1.4354	1.5218
HAHO	1.0443	0	1.3580	2.4742	1.5002	1.4345	1.5257
HBHU	1.2751	1.3580	0	2.5179	1.4281	1.3648	1.5242
CCPG	2.4541	2.4742	2.5179	0	2.3658	2.5065	2.6229
LEGH	1.4264	1.5002	1.4281	2.3658	0	1.4117	1.5876
MYWHP	1.4354	1.4345	1.3648	2.5065	1.4117	0	1.6261
FGFH	1.5218	1.5257	1.5242	2.6229	1.5876	1.6261	0

Besides the digital coding methods, topology structures representations are other important methods reflecting the secondary or the tertiary structures of proteins. TOPS diagrams are simple but useful descriptions of the structural topology of proteins. The objects of TOPS diagrams are the secondary structure elements (SSEs) in which each α-helix is a circle and each β-strand is a triangle. Each SSE in a TOPS diagram is associated with a direction represented with up or down, which is relative to the protein fold. The SSEs is usually represented by characters

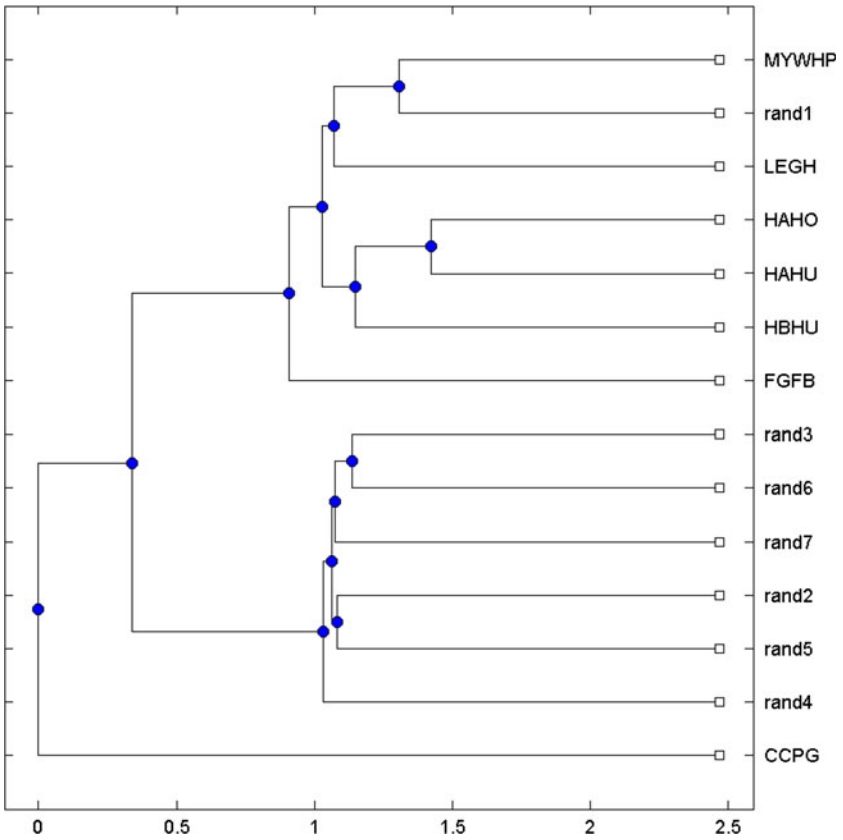

Fig. 1 The cluster tree using UPGMA method for the seven proteins and the seven random sequences. The HAHU, HAHO, and HBHU are in the same group, and MYWHP and LEGH in another group. All but one random sequence are in the same group which is different from the group consisting of five similar proteins

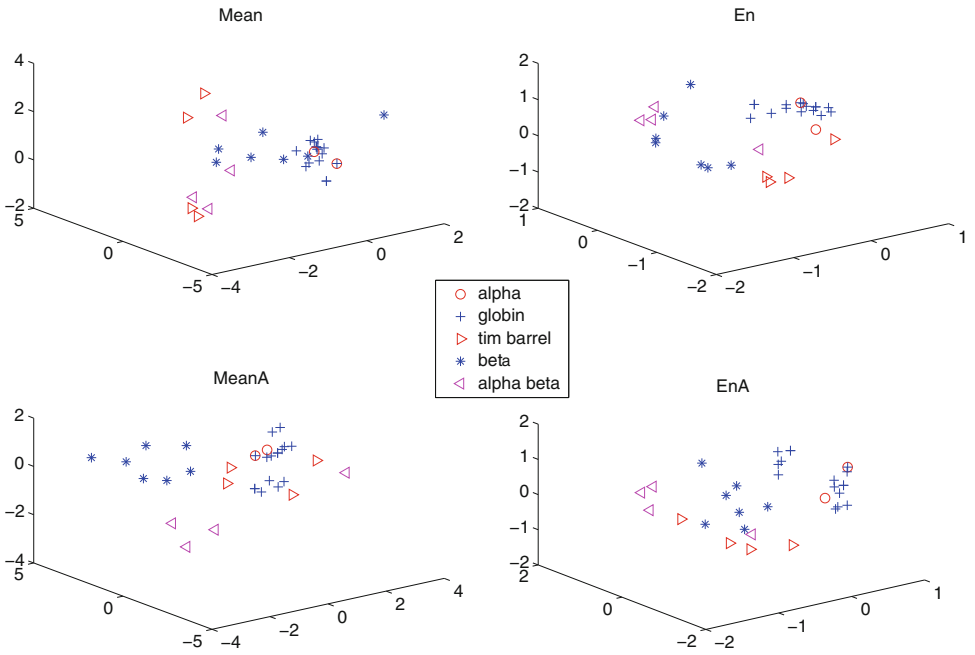

Fig. 2 The multidimensional scaling of the distances D_M, D_E with TOPS strings (*Mean* and *En*) and with ATOPS strings (*MeanA* and *EnA*). The distance D_E with ATOPS strings performs better than the others. The globin and alpha proteins are clearly separated from the others

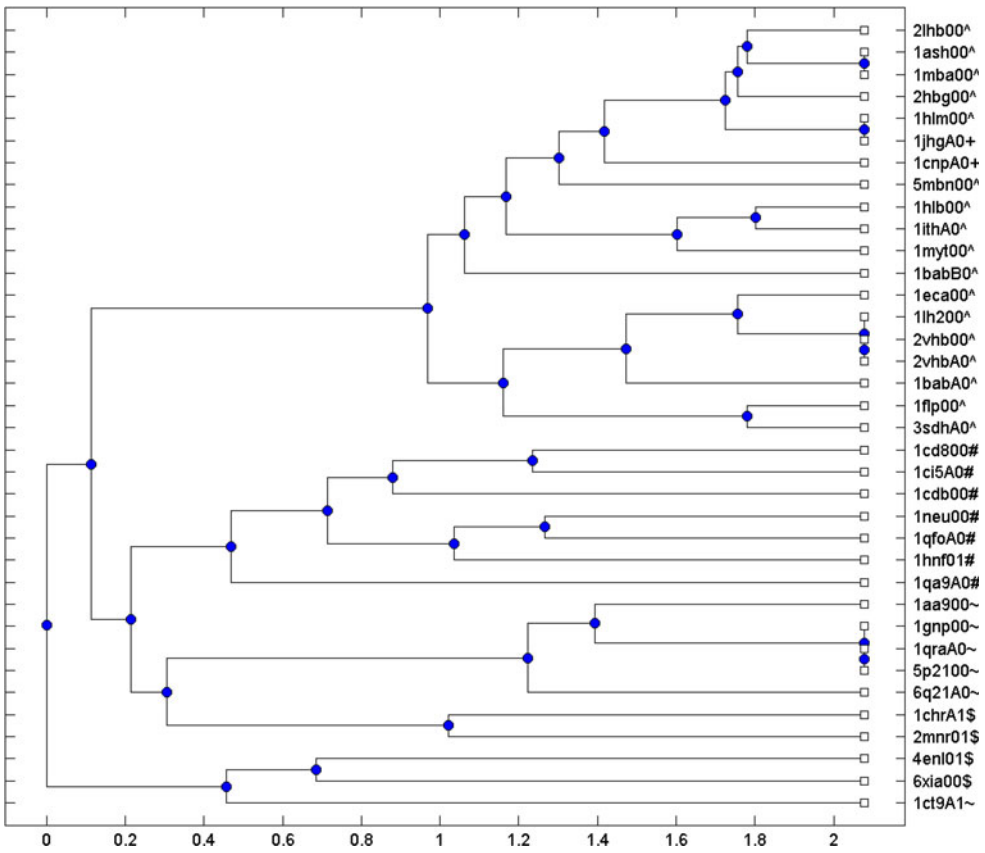

Fig. 3 The cluster tree of Chew–Kedem data set constructed by distance matrix D_E with UPGMA method

belonging to the alphabet $\{h, H, e, E\}$, where h is for the 'down' helix, H is for the 'up' helix, e is for the 'down' strand and E is for the 'up' strand. More details about the TOPS diagrams are described in Ferragina et al. (2007), Chew and Kedem (2003) and Liu and Wang (2008). However, it is hard for BWSD to compare the difference between h and H with the difference between h and E. To represent the direction, we use the '0' and '1' rather than the capital letters. To show the difference between the direction and the class of the SSE, we write the letters double. According to those, we change the alphabet $\{h, H, e, E\}$ to $\{hh1, hh0, ee1, ee0\}$. The strings of SSEs based on set $\{hh1, hh0, ee1, ee0\}$ are called adjusted TOPS (ATOPS) strings. For example: the 1hnf01's TOPS string is *NeEehEHeEeHEeC* while its ATOPS is *Nee*1*ee*0*ee*1*hh*1*ee*0*hh*0*ee*1*ee*0*ee*1*hh*0*ee*0*ee*1*C*.

Experiments and results

Experiment 1: the comparison of the protein primary sequences

To test the performance of our proposed approach with protein sequences, we selected the same protein data sets which were studied by Trad et al. (2002) and Pham (2007).

Table 7 Transferrin sequences, sources, and accession numbers

Sequence Name	Species	Accession No.
Human TF	*Homo sapien*	S95936
Rabbit TF	*Oryctolagus coniculus*	X58533
Rat TF	*Rattus norvegicus*	D38380
Cow TF	*Bos taurus*	U02564
Buffalo LF	*Bubalus arnee*	AJ005203
Cow LF	*Bos taurus*	X57084
Goat LF	*Capra hircus*	X78902
Camel LF	*Camelus dromedaries*	AJ131674
Pig LF	*Sus scrofa*	M92089
Human LF	*H. sapiens*	NM_002343
Mouse LF	*Mus musculus*	NM_008522
Possum TF	*Trichosurus vulpecula*	AF092510
Frog TF	*Xenopus laevis*	X54530
Japanese flounder TF	*Paralichthys olivaceus*	D88801
Atlantic salmon TF	*Salmo salar*	L20313
Brown trout TF	*Salmo trutta*	D89091
Lake trout TF	*Salvelinus namaycush*	D89090
Brook trout TF	*Salvelinus fontinalis*	D89089
Japanese char TF	*Salvelinus pluvius*	D89088
Chinook salmon TF	*Oncorhynchus tshawytscha*	AH008271
Coho salmon TF	*Oncorhynchus hisutch*	D89084
Sockeye salmon TF	*Oncorhynchus nerka*	D89085
Rainbow trout TF	*Oncorhynchus mykiss*	D89083
Amago salmon TF	*Oncorhynchus masou*	D89086

The protein sequences with their GenBank identities are shown in Table 2. In terms of structural classification, human α hemoglobin (HAHU), horse α hemoglobin (HAHO), human β hemoglobin (HBHU), Lupine leghemoglobin (LEGH) and sperm whale myoglobin (MYWHP) belong to the same super family, and others to a different super family. Based on the protein structure, HAHU, HAHO, HBHU, LEGH, MYWHP are more similar than others. HAHU and HAHO are orthologous sequences. Orthologous proteins are found in two or more organisms that have very high similarity and almost have similar three-dimensional structure, domain structure, and biological function (Pham 2007). HAHU and HBHU are paralogous sequences. Those sequences have about 43% of identical residues.

Tables 5 and 6, respectively, show the similarities of D_M and D_E of the seven protein sequences. Tables 3 and 4 show the result of the method based on wavelet (Trad et al. 2002) and the method based on spectral distortion measure (Pham 2007) with the same seven protein sequences. For the wavelet-based approach, the cross-correlation of HAHU and HAHO is 5S (highly correlated); between HAHU and HBHU the cross-correlation is 1S1W3N; and the cross-correlation of LEGH and MYWHP against HAHU are 2W3N. Those sequences belong to the same protein family. Even though LEGH has only 14% identical residues with HAHU, they are distantly related. Too weak correlations are detected in two different resolutions by the wavelet-based method. While if 0.005 is chosen to the cutoff value, the spectral distortion measure can detect HAHU, HAHO, HBHU and MYWHP as related proteins. The similar results can be found by our method with which the cutoff value for D_E and D_M are 1.45 and 0.55, respectively. Further, seven random sequences with the same size of the seven proteins mentioned above are generated to detect the impact of the random sequence. Figure 1, which is the cluster tree using UPGMA method for the seven proteins and the seven random sequences, shows that the HAHU, HAHO, HBHU are in the same group and MYWHP, LEGH in another group. All but one random sequence is in the same group which is different from the group consisting of five similar proteins. That's to say, the

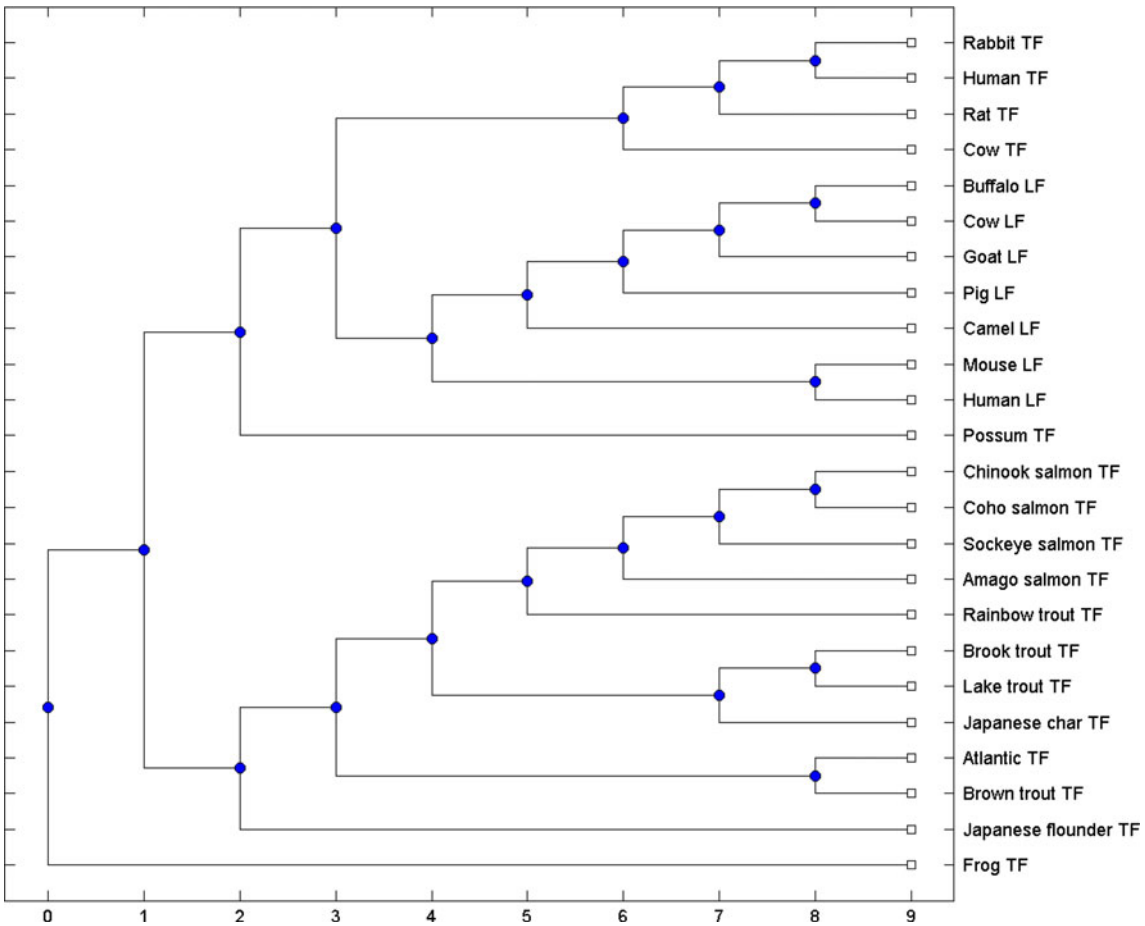

Fig. 4 The phylogenetic tree of transferrin sequences generated by Ford (2001)

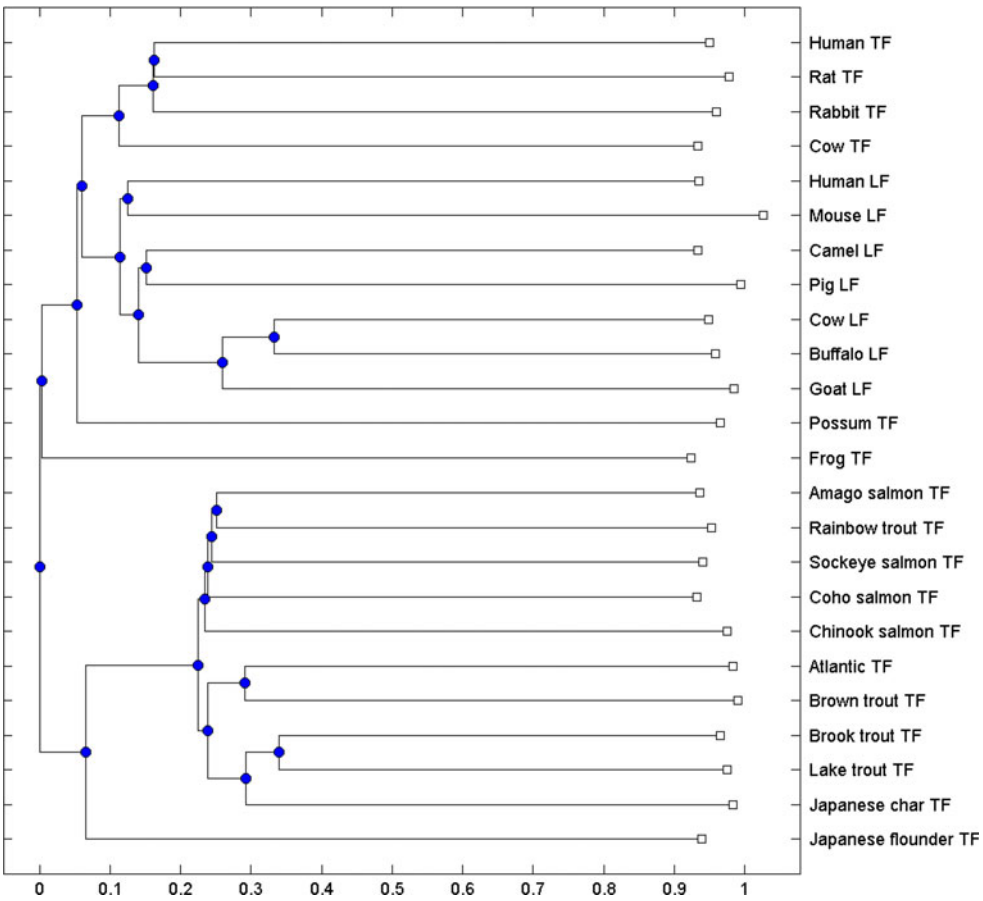

Fig. 5 The phylogenetic tree of transferrin sequences generated by the distance D_M

random sequences make little trouble in comparing the protein sequences using our approach.

Experiment 2: the comparison of the structure of the proteins based on adjusted TOPS strings

The Chew–Kedem data set, which is introduced by Chew and Kedem (2003) and well studied by Ferragina et al. (2007), Liu and Wang (2008) and Guo and Wang (2008), is selected to test the performance of the structure comparison on proteins. The proteins are as follows:

globin: 1eca00, 5mbn00, 1hlb00, 1hlm00, 1babA0, 1babB0, 1ithA0, 1mba00, 2hbg00, 2lhb00, 3sdhA0, 1ash00, 1flp00, 1myt00, 1lh200, 2vhbA0, 2vhb00;
alpha–beta: 1aa900, 1gnp00, 6q21A0, 1ct9A1, 1qraA0, 5p2100;
tim barrel: 6xia00, 2mnr01, 1chrA1, 4enl01;
beta: 1cd800, 1ci5A0, 1qa9A0, 1cdb00, 1neu00, 1qfoA0, 1hnf01;
alpha: 1cnpA0, 1jhgA0.

The TOPS strings for 36 protein chains can be found in Liu and Wang (2008). Using the TOPS strings, we can obtain the ATOPS strings of the 36 proteins. Then we compute the distance of each pair of the proteins by BWSD. When the distance matrices of the proteins are obtained, the multidimensional scaling method (Borg and Groenen 1997; Shepard 1966) is used to analyze those distance matrices. Multidimensional scaling (MDS) allows you to visualize how near points are to each other for many kinds of distance or dissimilarity metrics and can produce a representation of your data in a small number of dimensions. MDS does not require raw data but only a matrix of pairwise distances or dissimilarities. Figure 2 shows the results of the distances D_M, D_E with TOPS strings and with ATOPS strings. The distance D_E with ATOPS strings performs better than the others. The globin and alpha proteins are separated from the others clearly. Although there are some mixes between the globin and tim barrel proteins in the result of D_M with ATOPS strings, the beta proteins are separated more clearly than other methods, and the distance between 1cnpA0 and 1jhgA0 are the smallest in all the methods. Besides those results, there is a phenomenon that the globin proteins are divided into two groups clearly by the D_E with ATOPS strings. The reason for this phenomenon may be worth further considering. Those results are reappeared by the cluster tree constructed by

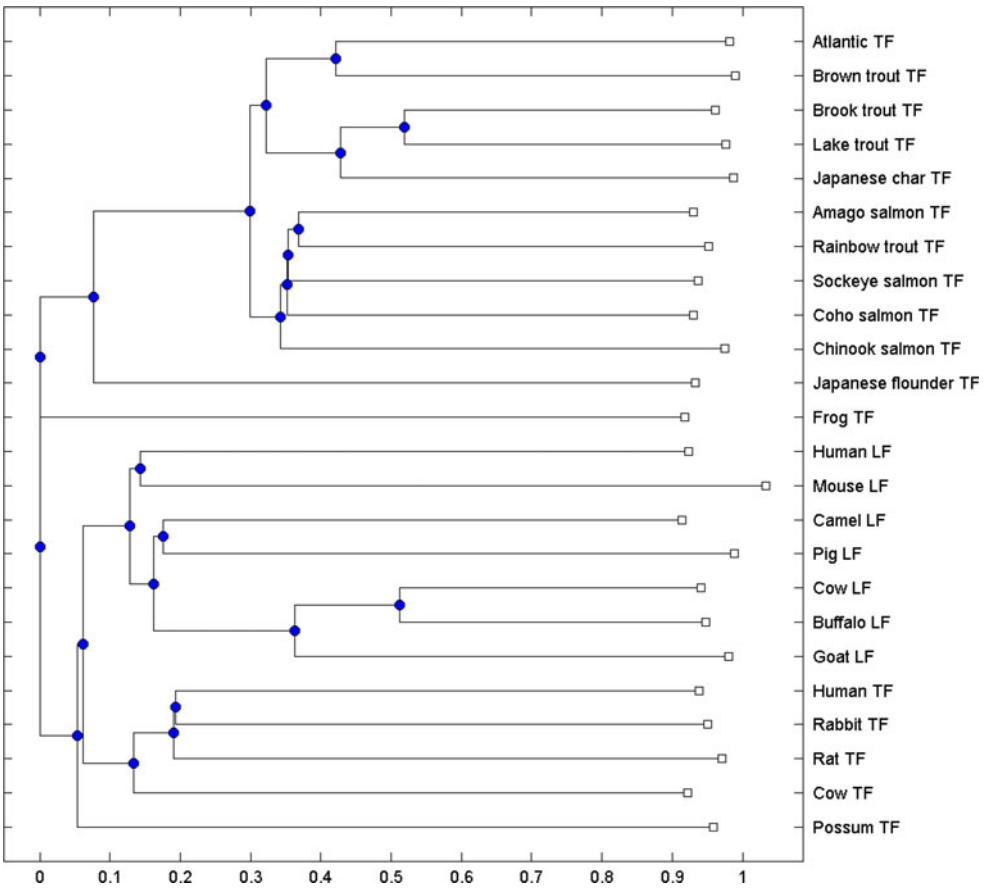

Fig. 6 The phylogenetic tree of transferrin sequences generated by the distance D_E

Table 8 The comparison of phylogenetic trees by symmetric difference distance

	D_MUP	D_MNJ	D_EUP	D_ENJ
Non-code	28	22	22	22
Code	12	12	12	10

distance matrix D_E with UPGMA method and illustrated in Fig. 3.

Experiment 3: the phylogenetic analysis of transferrin proteins

The transferrin sequences from 24 vertebrates, which are well studied by Ford (2001) (The accession numbers are listed in Table 7 and the topology structure of the phylogenetic tree is illustrated in Fig. 4), are selected to test the performance of our proposed approach on constructing the phylogenetic tree. After converting the protein sequences to the digital codes, we can obtain the BWSDs of all species pairs through the bwt algorithm. Once the distance matrices named D_M and D_E are obtained from BWSD, it is straightforward to generate a phylogenetic tree using NJ method or UPGMA method. The widely known symmetric difference distance of Robinson and Foulds (1981) is employed to compare the phylogenetic trees. The symmetric difference is simply a count of how many partitions there are, among the two trees, which are on one tree and not on the other. When the two trees are fully resolved bifurcating trees, their symmetric distance must be an even number; it can range from 0 to twice the number of internal branches, which for n species is $4n - 10$ (for three species or more).

The phylogenetic trees illustrated in Figs. 5 and 6 are constructed by D_M and D_E using NJ method. Both are substantially consistent with the tree constructed by Ford (2001). Further, considering the impact of the order relation on the alphabet set A, we construct the phylogenetic trees based on D_M and D_E computed without converting the protein sequences to the digital codes and compare these phylogenetic trees, using symmetric difference distance, with the tree constructed by Ford (2001). Table 8 lists the distance values. It can be found that (1) the trees constructed after converting the proteins into digital codes perform better than the trees constructed with the primary sequence without coding, (2) the trees constructed by NJ method perform better than the trees constructed by the UPGMA method, and (3) the trees constructed by the

distance measure D_E perform better than the trees constructed by the distance measure D_M.

The widespread occurrence and diverse roles of the family of transferrin proteins drive investigations into their evolution and function. Their binding and transport of hydrolysis prone Fe(III) are well studied. The correct construction of the phylogenetic trees shows that the hydrophobicity plays an important role in the evolution of the transferrin proteins.

Hence the Burrows–Wheeler similarity distribution is valid on similarity analysis. In the field of phylogenetic analysis, the distance measures D_M and D_E are efficient.

Conclusions

A novel alignment-free method for protein sequence comparison is proposed in this work. We use the BWSD to describe the similarity of the two proteins. The BWSD needs an appropriate representation which can reflect the properties of the proteins better. Hence, when we consider the primary sequences of the proteins, we use the digital codes which can reflect the amino acid's chemical physical properties to represent the protein sequences and when we consider the structure of the proteins, we use the ATOPS to describe the secondary structure of the proteins in this paper. Some experiments have shown that the approach proposed in this paper is a powerful and useful tool for the comparison of proteins. Besides the representation methods proposed in this paper, there are some other methods representing the proteins to reflect the properties we are interested in. If the BWSD is employed, the representations should match the BWSD which is a tool to describe the similarity of two strings.

Acknowledgments We would like to thank the reviewers for their useful and critical comments, all of which have greatly improved the quality of the paper. This work is supported by the National Natural Science Foundation of China (Grant No. 10871219) and by the National Science Foundation of China (Grant No. 10801026).

References

Blaisdell B (1986) A measure of similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci 83:5155–5159

Blaisdell B (1989) Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarities of natural sequences. J Mol Evol 29:526–537

Borg I, Groenen P (1997) Modern multidimensional scaling. Springer, New York

Burrows M, Wheeler D (1994) A block sorting data compression algorithm, Digital SRC Research Report

Cao Z, Liao B, Li R (2008) A group of 3D graphical representation of DNA sequences based on dual nucleotides. Inter J Quant Chem 108:1485–1490

Chen X, Francia B, Li M (2004) Shared information and program plagiarism detection. IEEE Trans Inf Theory 50(7):1545–1551

Cheng F, Shen J, Xu X, Luo X, Chen K, Shen X, Jiang H (2009) Interaction models of a series of oxadiazole-substituted alpha-isopropoxy phenylpropanoic acids against PPARalpha and PPARgamma: molecular modeling and comparative molecular similarity indices analysis studies. Protein Pept Lett 16:150–162

Chew L, Kedem K (2003) Finding the consensus shape for a protein family. Algorithmica 38(1):115–129

Chou KC (2004) Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134

Chou KC, Shen HB (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162

Cilibrasi R, Vitányi P, de Wolf R (2004) Algorithmic clustering of music based on string compression. Comput Music J 28(4): 49–67

Cristea P (2001) Independent componet analysis for genetic signals. In: SPIE conference BIOS 2001—international biomedical optics symposium, pp 20–26

Dai Q, Yang Y, Wang T (2008) Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. Bioinformatics 24(20):2296–2302

Dou Y, Zheng X, Wang J (2010) Several appropriate background distributions for entropy-based protein sequence conservation measures. J Theor Biol 262:317–322

Dou Y, Zheng X, Wang J (2009) Prediction of catalytic residues using the variation of stereochemical properties. Protein J 28:29–33

Feng J, Wang T (2008a) Characterization of protein primary sequences based on partial ordering. J Theor Biol 254(4): 752–755

Feng J, Wang T (2008b) Condensed representations of protein secondary structure sequences and their application. J Biomol Struct Dyn 25:621–628

Ford M (2001) Molecular evolution of transferrin: evidence for positive selection in salmonids. Mol Biol Evol 18:639–647

Ferragina P, Giancarlo R, Greco V, Manzini G, Valiente G (2007) Compression-based classification of biological sequences and structures via the universal similarity metric: experimental assessment. BMC Bioinformatics 8:252–272

Guo Y, Wang T (2008) A new method to analyze the similarity of protein structure using TOPS representations. J Biomol Struct Dyn 26:367–374

Helden JV (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20:399–406

Jia C, Liu T, Zhang X, Fu H, Yang Q (2009) Alignment-free comparison of protein sequences based on reduced amino acids alphabets. J Biomol Struct Dyn 26:763–770

Kantorovitz M, Robinson G, Sinha S (2007) A statistical method for alignment free comparison of regulatory sequences. Bioinformatics 23:i249–i255

Li M, Badger J, Chen X, Kwong S, Kearney P, Zhang H (2001) An information based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17:149–154

Li M, Chen X, Li X, Ma B, Vitányi P (2004) The similarity metric. IEEE Trans Inf Theory 12(5):3250–3264

Li M, Vitányi P (1997) An introduction to Kolmogorov complexity and its applications. Springer, Berlin

Liao B, Shan X, Zhu W, Li R (2006) Phylogenetic tree construction based on 2D graphical representation. Chem Phys Lett 422:282–288

Lin J (1991) Divergence measures based on the shannon entropy. IEEE Trans Inf Theory 37:145–151

Liu L, Wang T (2007) Novel characterization of the folding of proteins. Int J Quantum Chem 107:1970–1974

Liu L, Wang T (2008) Comparison of TOPS strings based on LZ complexity. J Theor Biol 251:159–166

Liu Y, Yang Y, Wang T (2007) Characteristic distribution of L-tuple for DNA primary sequence. J Biomol Struct Dyn 25:85–92

Mantaci S, Restivo A, Sciortino M (2003) Burrows-Wheeler transform and sturmian words. Inf Process Lett 86:241–246

Mantaci S, Restivo A, Rosone G, Sciortino M (2007) An extension of the burrows wheeler transform. Theor Comput Sci 387:298–312

Mantaci S, Restivo A, Sciortino M (2008) Distance measures for biological sequences: some recent approaches. Int J Approx Reason 47:1–18

Milligan G, Cooper M (1986) A study of the comparability of external criteria for hierarchical cluster analysis. Multivar Behav Res 21:441–458

Nandy A, Ghosh A, Nandy P (2009) Numerical characterization of protein sequences and application to voltage-gated sodium channel α subunit phylogeny. In Silico Biol 9:77–88

Otu H, Sayood K (2003) A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19(16):2122–2130

Pham T (2007) Spectral distortion measures for biological sequence comparisons and database searching. Pattern Recogn 40:516–529

Pham T, Zuegg J (2004) A probabilistic measure for alignment-free sequence comparison. Bioinformatics 20:3455–3461

Randić M (2007) 2-D graphical representation of proteins based on physico-chemical properties of amino acids. Chem Phys Lett 440:291–295

Randić M, Butina D, Zupan J (2006) Novel 2-D graphical representation of proteins. Chem Phys Lett 419:528–532

Robinson D, Foulds L (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147

Shepard R (1966) Metric structure in ordinal data. J Math Psych 3:287–315

Trad C, Fang Q, Cosic I (2002) Protein sequence comparison based on the wavelet transform approach. Protein Eng 15:193–203

Vinga S, Almeida J (2003) Alignment-free sequence comparison—a review. Bioinformatics 19(4):513–523

Xiao X, Chou K (2007) Digital coding of amino acids based on hydrophobic index. Prot Pept Lett 14:871–875

Yang L, Zhang X, Wang T (2009) The Burrows-Wheeler similarity distribution between biological sequences based on Burrows-Wheeler transform. J Theor Biol 262:724–749

Zhang C, Zhang R (2000) S curve, a graphic representation of protein secondary structure sequence and its applications. Biopolymers 53:539–549

Zhang S, Wang T (2010) Phylogenetic analysis of protein sequences based on conditional LZ complexity. MATCH Commun Math Comput Chem 63(3)

Zhang S, Yang L, Wang T (2009) Use of information discrepancy measure to compare protein secondary structures. J Mol Struct Theochem 909:102–106