Proceedings of the 5th WSEAS Int. Conf. on SIMULATION, MODELING AND OPTIMIZATION, Corfu, Greece, August 17-19, 2005 (pp457-462)

A new method to construct phylogenetic tree from proteins

Na Liu [1,2*], Tianming Wang[2,3]

[1] *Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China*
[2] *College Advanced Science and Technology, Dalian University of Technology, Dalian 116024, CHINA*
[3] *Department of Mathematics , Hainan Normal University, Haikou 571158, CHINA*
liunasophia@163.com , wangtm@dlut.edu.cn

Abstract: A phylogenetic analysis of a family of related biological sequences is to determinate how the family might have been derived during evolution. Current versions of phylogenetic analysis programs provide three main methods for phylogenetic analysis —parsimony, distance, and maximum likelihood methods—and also include many types of evolutionary models for sequence variation. Most of them make use of multiple alignment of sequences, which doesn't work for all types of data. Here we propose a new method for phylogenetic tree construction from proteins. It is based on characteristic sequence of protein. The proposed method doesn't require sequence alignment and complex algorithm and has reasonably constructed phylogenetic tree for real data set.

Key Words: phylogenetic tree, LZ complexity, distance measure, characteristic sequence, SMC proteins, MukB proteins, Rad50 proteins.

1 Introduction

With the accomplishment of Human Genome Project, the list of biological sequences is exploding. It is important to analyze their evolutionary history of different organisms or biological sequences. In present, there are three main methods for phylogenetic analysis: parsimony, distance, and maximum likelihood methods. They either calculate a matrix representing the distance between each pair of sequences or evaluate the fitness of different topologies [1]. All of these methods require a multiple alignment of the sequences and assume some sort of an evolutionary model. Multiple alignment becomes misleading due to gene rearrangement, inversion, transportation and translocation when the lengths of sequences are unequal. To overcome these problems, several distance measures using rearrangement, recombination, break point, comparative mapping and gene order have been studied for phylogeny [2-4] . These methods are computationally expensive. Snel *et al* [5] proposed gene content as a distance measure, which performs unefficiently when the gene content of the organisms are very similar. Later on, complexity was introduced into phylogenetic analysis. for example, Chen *et al* [6] and Li *et al* [7] defined the distance as $1-[K(S)-K(S|Q)]/K(SQ)$, where $K(S)$ is the kolmogorov complexity of S. It has been pointed out that kolmogorov complexity is an algorithmic measure of information that has a theoretical limit. Recently, Otu, H.H *et al* [8] described a new sequence distance measure for phylogenetic analysis and used it to successfully construct phylogenetic trees for real and simulated DNA data sets.

Motivated by the work of Otu, H.H *et al*, in this paper we propose a new method to construct phylogenetic tree from proteins. It is based on characteristic sequence of protein, which is derived according to the chemical properties of 20 amino acids. By this method, we construct the phylogenetic tree for 40 protein sequences from SMC, MukB, and Rad50 proteins, which is basically in accordance with the opinion of Cobbe, N and Heck, M.M.S.

*Fax: +(86)411-84706100.
E-mail addresses: liunasophia@163.com (Liu,N.), wangtm@dlut.edu.cn (Wang,T.M.)

2 Materials and method

2.1 The LZ complexity

Let S, Q and R be sequences over a finite alphabet Λ, $L(S)$ be the length of S, $S(i)$ be the ith element of S and $S(i,j)$ be the subsequence of S that starts at position i and ends at position j. Note that $S(i,j) = \emptyset$, for $i > j$. The contatenation of Q and R forms a new sequence $S = QR$, where Q is called a prefix of S, and S is called an extension of Q if there exists an integer i such that $Q = S(1,i)$.

An extension $S = QR$ of Q is reproducible from Q denoted by $Q \rightarrow S$, if there exists an integer $p \leq L(Q)$ such that $R(k) = S(p+k-1)$, for $k = 1, 2, .., L(R)$. For example: $ADDEF \rightarrow ADDEFDEF$ with $p = 3$. A non-null sequence S is producible from its prefix $S(1,j)$, denoted by $S(1,j) \Rightarrow S$, if $S(1,j) \rightarrow S(1, L(S)-1)$. For example: $SCCEF \Rightarrow SCCEFCCEG$ with $p = 2$.

Any non-null sequence S can be built from a production process by iterative self-deleting-building process where at the ith step $S(1,h_{i-1}) \Rightarrow S(1,h_i), \emptyset = S(1,0) \Rightarrow S(1,1)$. An m-step production process of S leads to a parsing of S into $H(S) = S(1,h_1) \bullet S(h_1+1,h_2) \bullet ..... \bullet S(h_{m-1}+1,h_m)$, which is called the history of S, and $H_i(S) = S(h_{i-1}+1,h_i)$ is called the ith component of $H(S)$.

A component $H_i(S)$ and the corresponding production step $S(1,h_{i-1}) \Rightarrow S(1,h_i)$ are called exhaustive if $S(1,h_{i-1}) \rightarrow S(1,h_i)$ is not true. A history is called exhaustive if each of its components (with a possible exception of the last one) is exhaustive. What's more important, the exhaustive history of any non-null sequence is unique. For example, for the sequence $S = PPCGAGGPCGGA$, its exhaustive history is $EH(S) = P \bullet PC \bullet G \bullet A \bullet GG \bullet PCGG \bullet A$.

Let $c(S)$ be the number of components in the exhaustive history of S. It is the least possible number of steps needed to generate S according to the rules of production process introduced by Lempel, A. *et al* [9]. Furthermore, the production process here is an example of a class of parsing rules, so $c(S)$ is an important complexity indicator.

2.2 Distance measures

According to Lempel, A *et al*, for any given sequences Q and S, $c(QS) \leq c(Q) + c(S)$ always remains valid. This formula shows that the steps required to extend Q to QS are always less than the steps required to build S from $\emptyset$. On the basis of their idea that the more similar the sequence S is to sequence Q, the smaller $c(QS) - c(Q)$ should be, Otu, H.H and Sayood, K defined five measures for describing the closeness degree between two sequences. For more information, please refer to [8]. They called them distance measures by regarding $D(S,S) = 0$ even if $D(S,S) \neq 0$ in theory. In other words, they regarded such non-zero number as an error. So each measure defined by them is an approximate distance measure in that a distance metric should satisfy the identity: $D(S,Q) \geq 0$, where the equality is satisfied iff $S = Q$. These five measures have been used to successfully construct phylogenetic tree using whole mtDNA sequences.

2.3 The characteristic sequence of protein

Proteins are chains of 20 amino acids joined by peptide bonds. The 20 amino acids found in proteins can be grouped according to the chemistry of their R groups as in [1]: amino acids A, V, F, P, M, I, L belong to the Hydrophobic chemical group; amino acids D, E, K, R belong to Charged chemical group; amino acids S, T, Y, H, C, N, Q, W belong to Polar chemical group; amino acid G belongs to Glycine chemical group.

Then for any protein sequence, we will transform it into a new sequence defined over alphabet

$\{H, C, P, G\}$. The rule is as follows:

$$R(S(i)) = \begin{cases} H\,, & S(i) = A, V, F, P, M, I, L\,, \\ C, & S(i) = D, E, K, R\,, \\ P, & S(i) = S, T, Y, H, C, N, Q, W\,, \\ G, & S(i) = G\,, \end{cases} \tag{1}$$

For example, for protein sequence $S = VFFPDETGTGSYHMRWGSTQQCQVFEGLDEQQ$, the transformation result is $T(S) = HHHHCCPGPGPPPHCPGPPPPPPHHCGHCCPP$.

This newly-got sequence describes the original protein sequence from the chemistry aspect of the 20 amino acids and hence can be regarded as a coarse-grained description of protein sequence. So we call it the characteristic sequence of protein. In the whole process of constructing phylogenetic tree, we will substitute the corresponding characteristic sequences for all proteins.

2.4 The calculation of the distance matrix.

Provided n protein sequences $Q_1, Q_2, \ldots\ldots, Q_n$, which are under the phylogeny analysis, we first transform them into their characteristic sequences denoted by $S_1, S_2, \ldots\ldots, S_n$. Then with the aid of computer, we make pair-contatenation operation on these characteristic sequences and obtain their exhaustive histories by parsing the sequences using the production rules described above and in [9]. That means we will get the number of components in each exhaustive history: $C(S_1), C(S_2), \ldots\ldots, C(S_n), C(S_1S_1), \ldots\ldots, C(S_1S_n), \ldots\ldots, C(S_nS_1), \ldots\ldots, C(S_nS_n)$.

Here we use the

$$d(S, Q) = \frac{(c(SQ) - c(S)) + (c(QS) - c(Q))}{c(SQ) + c(QS)} \tag{2}$$

as the distance measure to construct the phylogenetic tree from proteins, twice of which is identical to the fifth formula defined by Otu, H.H *et al.* Even more, the adoption of (2) can result in much smaller error in that the smaller the coefficient of (2), the smaller the error will be. The choice of certain appropriate coefficient depends on the biologist's interest because, on a whole, it has little influence on the result of phylogenetic tree. Now we may obtain an approximate distance matrix by regarding all the entries on the diagonal as zeros even if they may be not zeros theoretically, *i.e.* we regard them as error terms as Otu, H.H. *et al* have done. Note that this distance matrix reflects the similarity of the n original protein sequences from their characteristic sequences. Hence we use it to construct the phylogenetic tree of the n proteins. Now what's left to be done is only to transform this matrix into phylogenetic tree like other matrices by choosing certain program such as Neighbor Joining, UPGMA etc.

3 Results and discussion

In order to test the validity and efficiency of our method, we will look at how well the result generated by our method agrees with the existing phylogenies. Our testing data contains all six SMC subfamilies and its related MukB proteins and Rad50 proteins from Eukaryotes and Prokaryotes. They are retrieved from NCBI database. The SMC proteins are found in nearly all living organisms, where they play crucial roles in mitotic chromosome dynamics, regulatoin of gene expression and DNA repair. These proteins are a highly conserved and ubiquitous family and the function of these

proteins in different aspects of chromosomal behaviour is conserved in eukaryotes and prokaryotes. We use the UPGMA option in Neighbour program to construct the phylogenetic tree. The final phylogenetic tree of these protein sequences is viewed using TreeView program, shown in Figure1.

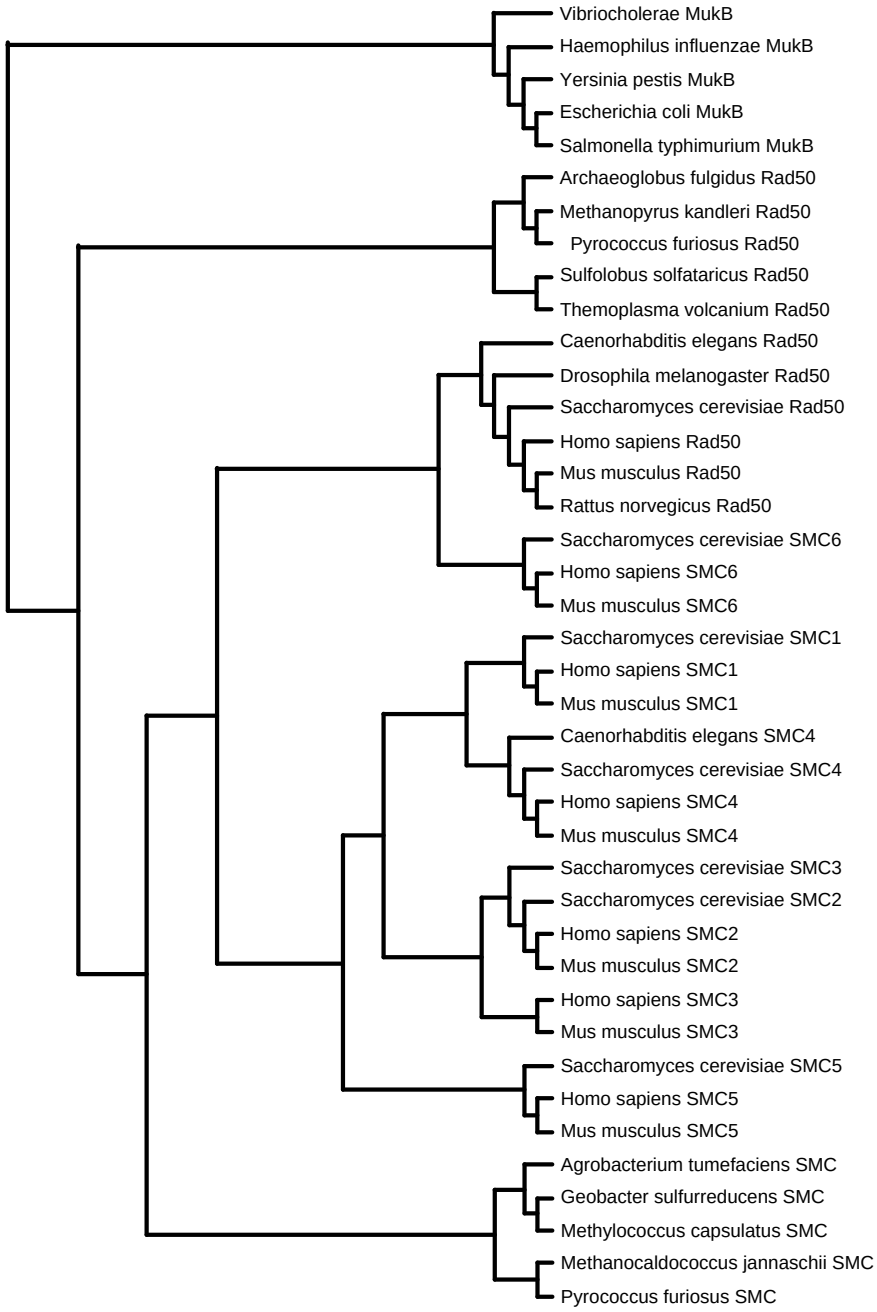

Figure1. The phylogenetic tree of 40 proteins by our method

We observe that all SMC proteins are grouped together with SMC1 and SMC4, SMC2 and SMC3 more closely. All MukB proteins , all Archaeal Rad50 proteins and all Eukaryotic Rad50 proteins are closely grouped, respectively. These coincide with the result that Cobbe, N *et al* have got by maximum liklihood method [10, 11]. Note that, our tree supports the opinion of Cobbe, N *et al* that SMC proteins should be more closely related to Rad50 proteins than MukB proteins. Of course, there exists little difference from the result obtained by Cobbe, N *et al.* For example, Archaeal Rad50 and Eukaryotic Rad50 subfamilies are not grouped very closely, instead Eukaryotic Rad50 family and SMC6 protein subfamily are grouped closely; SMC5 and SMC6 protein subfamilies are not grouped very closely like SMC2 and SMC3. Moreover, Saccharomyces cerevisiae SMC3 and Saccharomyces cerevisiae SMC2 are put in same branch. In fact, as pointed out, a number of different phylogenies for SMC proteins and its related proteins have been suggested to date, which differ in their methods of construction and resultant topology [10-14].So the precise relationships between them have been unclear. Ours is basically in agreement with what Cobbe, N *et al* have obtained.

4 Conclusion

In this paper, we propose a new method for constructing phylogenetic tree from proteins. It makes use of the characteristic sequences of proteins. This method is motivated by the idea of Otu, H.H and Sayood, K that associates the resulting number of steps to generate a sequence from a different sequence with the "closeness" between the two sequences. Unlike the traditional methods, we don't use the protein sequences directly, but use the transformation of them—characteristic sequences to study the phylogeny. That is to say, we analyze the phylogeny from the aspect of chemical properties of 20 amino acids comprising proteins. This method doesn't require sequence alignment and avoids to adopt appropriate model of amino acid replacement. Unequal sequence length or the relatively different positions of similar regions between sequences is not problematic.

The phylogenetic tree shows that the proposed method can reasonably construct phylogenetic tree from proteins. Finally, it is worth noting the role of the characteristic sequences of proteins. What does the reasonableness of the phylogenetic tree from proteins constructed by characteristic sequences imply? Perhaps, referring to the evolution of species, biologists can give us some heuristic explanation.

References

[1] Mount, D.W., *Bioinformatics: Sequence and genome analysis*, Cold Spring Harbor Laboratory Press, 2001.

[2] Kececioglu, J., Ravi, R., Of mice and men. Evolutionary distances. In *Proceedings of the 6th ACM-SIAM Symposium on Discrete Algorithms*, 1995, 604-613.

[3] Sankoff, D., Genome rearrangement with gene families, *Bioinformatics*, 15, 1999, 909-917.

[4] Sankoff, D., Blanchette, M., Multiple genome rearrangement and breakpoint phylogeny, *J. Comput. Biol,* 5, 1998, 555-570.

[5] Snel, B., Bork, P., Huynen, M.A., Genome phylogeny based on gene content. *Nat. Genet,* 21, 1999, 108-110.

[6] Chen, X., kwong, S., Li, M., A compression algorithm for DNA sequences and its applications in genome comparison. In Shamir, R., Miyano, S., Istrail, S., Pevzner, P., Waterman, M. (eds), In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB)*, ACM Press, Tokyo, Japan, 2000, 107-117.

[7] Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., Zhang, H., An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics,* 17, 2001, 149-154.

[8] Otu, H.H., sayood, K., A new sequence distance measure for phylogenetic tree construction, *Bioinformatics,* 25, 2003, 1364-1368.

[9] Lempel,A., Ziv,J., On the complexity of finite sequences, *IEEE T. Inform. Theory,* 22, 1976, 75-81.

[10] Cobbe, N., Heck, M.M.S., Review: SMCs in the world of chromosome biology—from prokaryotes to higher eukaryotes, *J. Struct. Biol,* 129, 2000, 123-143.

[11] Beasley, M., Xu, H., Warren, W., McKay, M., Conserved disruptions in the predicted coiled-coil domains of eukaryotic SMC complexes: implications for structure and function, *Genome Res,* 12, 2002,1201-1209.

[12] Melby, T.E., Ciampaglio, C.N., Briscoe, G., Erickson, H.P., The symmetrical structure of structral maintenance of chromosomes (SMC) and MukB proteins: long, antiparallel coiled coils, folded at a flexible hinge, *J.Cell biol,* 142, 1998, 1595-1604.

[13] Soppa, J., Prokaryotic structural maintenance of chromosomes (SMC) proteins: distribution, phylogeny, and comparison with MukBs and additional prokaryotic and eukaryotic coiled-coil proteins, *Gene,* 278, 2001, 253-264.

[14] Cobbe, N., Heck, M.M.S., *The evolution of SMC proteins: Phylogenetic analysis and structural implications,* MBE Advance Access, 2003.