Decomposition and Reconstruction of Protein Sequences: The Problem of Uniqueness and Factorizable Langauge

Xiaoli Shi
Institute of Botany, Academia Sinica, Beijing, China

Huimin Xie
Department of Mathematics, Suzhou University, Suzhou, China

Shuyu Zhang
Institute of Physics, Academia Sinica, Beijing, China

Bailin Hao*
T-Life Research Center, Fudan University, Shanghai, China,
and Institute of Theoretical Physics, Academia Sinica, Beijing, China

(Received 16 August 2006)

In our attempt to justify the CVTree approach of inferring phylogenetic relationship among bacteria from their complete genomes without using sequence alignment, we encountered the problem of the uniqueness of the reconstruction of a protein sequence from its constituent K-peptides, which has a natural relation to a well-understood problem in graph theory, namely, the number of Eulerian loops in a graph. The existence of finite state automata to recognize the uniqueness of a sequence reconstruction provides us with yet another application of factorizable language, which was elucidated at a previous Dynamics Days Asia Pacific meeting (DDAP1).

PACS numbers: 87.10.+e, 87.15.+e, 87.14.Ee
Keywords: Proteins, Repeated peptides, Eulerian loops, Graph theory, Factorizable language

I. INTRODUCTION

In recent years we have developed a composition vector approach, abbreviated as CVTree [1,2], to infer the phylogenetic relationship of prokaryotes from their complete genomes. As bacterial genomes differ significantly in their sizes and gene contents, one has to give up sequence-alignment-based methods. Instead of using the primary protein sequences, we decompose all the protein sequences of a species into (overlapping) K-peptides and build a composition vector from the number of different K-peptides. An essential step in this approach is the subtraction of a random background by using a weak Markovian assumption, which has led to phylogenetic results comparable in details with the century-long effort of bacterial taxonomy. For a brief review for physicists see [3]. The justification of the CVTree [2] approach has inspired the problem of the uniqueness of the reconstruction of a protein sequence from its constituent K-peptides, which will be formulated later.

The composition vector approach may be put into a broader context. The composition of nucleotides in DNA sequences or the amino acid frequencies in protein sequences have been widely used in characterizing DNA or proteins. For example, the $g + c$ content or CpG islands have played an important role in gene-finding programs. However, this kind of study usually has been restricted to the frequency of single letters or short strings; *e.g.*, dinucleotide correlations in DNA sequences, *i.e.*, only the $K = 1$ or the $K = 2$ case has been more or less explored. The use of K-tuples takes into account short-range correlation within K letters and corresponds to using $(K - 1)$-th order Markov models to characterize the sequences. With K increasing, more and more species-specific and even gene-specific features may come to the surface. As the problem posed in this paper has a natural relation to the number of Eulerian loops in a graph, we commence with a brief summary of necessary graph theory notions.

*E-mail: hao@itp.ac.cn; http://www.itp.ac.cn/~hao/

II. NUMBER OF EULERIAN LOOPS IN AN EULER GRAPH

Eulerian paths and Euler graphs comprise a well-developed chapter of graph theory (see, *e.g.*, [4]). We collect a few definitions in order to fix our notation. Consider a connected directed graph made of a certain number of labeled nodes. A node i may be connected to a node j by a directed arc. From a starting node v_b, one may go through a number of arcs to reach an ending node v_f in such a way that each arc has been passed once and only once; such a path is then called an *Eulerian path.* If v_b and v_f coincide the path becomes an *Eulerian loop.* A graph in which an Eulerian loop exists is called an *Euler graph.* An Eulerian path may be modified to an Eulerian loop by drawing an auxiliary arc from v_f back to v_b. We only consider Euler graphs defined by an Eulerian loop.

From a given node, there may be d_{out} arcs going out to other nodes; d_{out} is called the outdegree (fan-out) of the node. There may be d_{in} arcs coming into a node, d_{in} being the indegree (fan-in) of the node. The condition for an undirected graph to be Eulerian was indicated by Euler in 1736. In our case of directed graphs, it may be formulated as

$$d_{\text{in}}(i) = d_{\text{out}}(i) \equiv d_i$$

for all nodes numbered in a certain way from $i = 1$ to m. The numbers d_i are simply called degrees. We put all degrees into a diagonal matrix:

$$M = \text{diag}(d_1, d_2, \cdots, d_m).$$

The connectivity of the nodes is described by an adjacent matrix $A = \{a_{ij}\}$, where a_{ij} is the number of arcs leading from node i to j. From the matrices M and A, one forms the Kirchhoff matrix

$$C = M - A.$$

The Kirchhoff matrix has the peculiar property that its elements along any row or column sum to zero:

$$\sum_i c_{ij} = 0, \ \sum_j c_{ij} = 0.$$

Furthermore, for any $m \times m$ Kirchhoff matrix all $(m-1)\times(m-1)$ minors are equal, and we denote this common minor by Δ.

A graph is called simple if between any pair of nodes, there are no parallel (repeated) arcs, and at all nodes, there are no rings; *i.e.*, $a_{ij} = 0$ or $1\ \forall i, j$ and $a_{ii} = 0\ \forall i$. The number R of Eulerian loops in a simple Euler graph is given by

The BEST Theorem [4] (BEST stands for N. G. de **B**rujin, T. van Aardenne-**E**hrenfest, C. A. B. **S**mith, and W. T. **T**utte):

$$R = \Delta \prod_i (d_i - 1)!. \tag{1}$$

The BEST formula in Eq. (1) gives the number of Eulerian loops in an Euler graph without specifying a starting node. If a node k is specified as the beginning (hence ending) of the loop, then the number of loops starting from k is [5]

$$R = \Delta d_k \prod_i (d_i - 1)!, \tag{2}$$

where d_k is the degree of the node k. In what follows, we consider only Eulerian loops with the starting node fixed.

For a general Euler graph, there may be arcs going out and coming into one and the same node (some $a_{ii} \neq 0$), as well as parallel arcs leading from node i to j ($a_{ij} > 1$). It is enough to put auxiliary nodes on each parallel arc and ring to make the graph simple. By applying elementary operations to the larger Kirchhoff matrix thus obtained, one can reduce it essentially to the original size with some $a_{ii} \neq 0$ and $a_{ij} > 1$. In accordance with the unlabeled nature of parallel arcs and rings, one must eliminate the redundancy in the counting result by dividing it by $a_{ij}!$. Thus, the BEST formula is modified to

$$R = \frac{\Delta d_k \prod_i (d_i - 1)!}{\prod_{ij} a_{ij}!}. \tag{3}$$

As $0! = 1! = 1$ Eq. (3) reduces to Eq. (2) for simple graphs. To the best of our knowledge, the modified BEST formula Eq. (3) first appeared in [6] where Eulerian loops from a fixed starting node were considered.

III. EULERIAN GRAPH FROM A PROTEIN SEQUENCE

We decompose a given protein sequence of length L into a set of $L - K + 1$ overlapping K-peptides by using a window of width K, sliding one letter at a time. Combining repeated peptides into one and recording their copy number, we get a collection $\{W_j^K, n_j\}_{j=1}^M$, where $M \leq L - K + 1$ is the number of different K-peptides.

Now, we pose the inverse problem. Given the collection $\{W_j^K, n_j\}_{j=1}^M$ obtained from the decomposition of a protein sequence, reconstruct all possible amino acid sequences subject to the following constraints:

1. Keep the start K-peptide unchanged, because most protein sequences start with methionine (M); even the tRNA for this initiation M is different from that for prolongation.
2. Use each W_j^K string n_j times until the given collection is used up.
3. The reconstructed sequence must be of the original length L.

Clearly, the inverse problem has at least one solution: one can always recover the original protein sequence. For

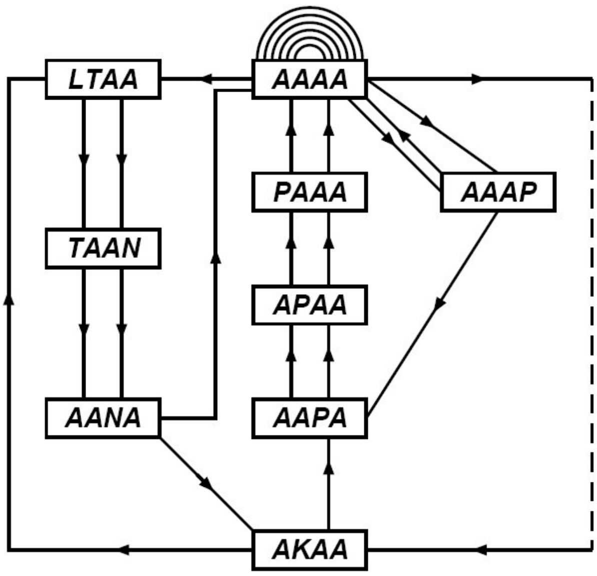

Fig. 1. An Euler graph derived from the protein ANPA_PSEAM sequence at $K = 5$.

some sequences, it may have multiple solutions. However, for K big enough, the solution must be unique, as evidenced by the extreme choice $K = L - 1$. We are concerned with how unique is the reconstruction for real proteins. The uniqueness of most of the natural protein sequences will speak in favor of the composition vector approach as it brings the collections of K-peptides closer to the primary protein sequences that are used in the traditional alignment-based methods.

In order to tell the number of reconstructed sequences, we first transform the original protein sequence into an Euler path by considering the two $(K-1)$ substrings of a K-peptide as two nodes and by drawing a directed arc to connect them. The same repeated $(K-1)$ substrings are treated as a single node with more than one incoming and outgoing arcs. Thus, we obtain a path from a beginning node v_b to an ending node v_f, where v_b is labelled by the left $(K-1)$-substring of the first K-peptide of the protein sequence and v_f is labelled by the right $(K-1)$-substring of the last K-peptide in the protein sequence. Generally speaking, $v_b \neq v_f$, and we can add an auxiliary node v_0 in between v_f and v_b and draw auxiliary arcs from v_f to v_0 and then from v_0 to v_b to make the path a closed loop. This Eulerian loop defines an Euler graph, and we are concerned with the number of different Eualerian loops in this graph with the node v_0 fixed. Since the degree of v_0 is always $d_0 = 1$, the modified BEST formula (3) takes the simpler form

$$R = \frac{\Delta \prod_i (d_i - 1)!}{\prod_{ij} a_{ij}!}. \tag{4}$$

Among natural proteins there are a few with $v_b = v_f$ that require some additional analysis. We skip that discussion and note that Eq. (4) works as well.

Take the protein database SWISS-PROT [7] entry ANPA_PSEAM as an example. This antifreeze protein A/B precursor of winter flounder has a short sequence of 82 amino acids. Some of its repeated segments are related to alanine-rich helices. The sequence reads

$$\begin{array}{lll} MALSLFTVGQ & LIFLFWTMRI & TEASPDPAAK \\ AAPAAAAAPA & AAAPDTADDA & AAAAALTAAN \\ AKAAAELTAA & NAAAAAAATA & RG \qquad . \end{array}$$

Consider the case $K = 5$. The first 5-peptide $MALSL$ gives rise to a transition from node $\boxed{MALS}$ to node $\boxed{ALSL}$. Shifting by one letter, from the next 5-peptide $ALSLF$, we get an arc from $\boxed{ALSL}$ to node $\boxed{LSLF}$, and so on, and so forth. Clearly, we get an Eulerian path whose nodes all have equal indegrees and outdegrees, except for the first and the last ones. Now, we draw an auxiliary arc from the last node $\boxed{TARG}$ to the first node $\boxed{MALS}$ to get a closed Eulerian loop. In general, one draws an auxiliary arc from the last node v_f to an auxiliary node v_0 and then another auxiliary arc from v_0 to the beginning node v_b to form a loop. We skip the special case $v_b = v_f$ for which a separate discussion is needed.

The Euler graph is defined by the above loop. In order to get the number of different Eulerian loops in this graph with v_0 being the fixed starting node, we have no need to generate a fully-fledged graph with all the distinct $(K-1)$-strings treated as nodes. Because we are interested only in the number of Eulerian loops, the graph can be simplified in several ways. For example, a series of consecutive nodes with $d_i = 1$ may be replaced by a single arc. In other words, only those strings in $\{W_j^{K-1}, n_j\}$ with $n_j \geq 2$ are used in drawing the graph. In our example the short list

$$\{AKAA, 2;\ AAPA, 2;\ APAA, 2;\ PAAA, 2;\ AAAA, 10;\ AAAP, 2;\ LTAA, 2;\ TAAN, 2;\ AANA, 2\}$$

leads to a small Euler graph with only 9 nodes (see Fig. 1).

The corresponding Kirchhoff matrix is:

$$C = \left\{ \begin{array}{ccccccccc} 2 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 4 & -2 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 \\ -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 2 \end{array} \right\}.$$

The minor $\Delta = 192$, and by using $R(K)$ to denote the number of reconstructions at a given K, we have $R(5) = \Delta 9!/6!2^6 = 1512$. For this protein $R(6) = 60$, $R(7) = 2$, and $R(8) = 1$.

We have written two programs: one exhaustively reconstructs amino acid sequences from a given protein, while the other implements the modified BEST formula in Eq. (4). The two programs yield identical results whenever comparable - we had to impose a cutoff to let the exhaustive enumeration program skip a protein when the number of reconstructions exceeded a preset value, say, 10 000. However, there is a third way to tackle the uniqueness problem.

IV. FINITE STATE AUTOMATON FOR CHECKING THE UNIQUENESS OF THE RECONSTRUCTION

Recently, there appeared a paper [8] studying the uniqueness of the reconstruction problem in an entirely different context. A theorem was proved in Ref. [8] that there exists a finite state automaton (FSA) that can judge whether a sequence has a unique reconstruction at a given K, but no explicit automaton was built. The existence of such a FSA becomes evident if we recollect the notion of a factorizable language that we discussed at DDAP1 [9].

Take the 20 amino acid symbols as an alphabet Σ. Denote by Σ^* the collection of all possible sequences over Σ, plus an empty string ϵ. In formal language theory, any subset $L \subset \Sigma^*$ may be defined as a language (see, *e.g.*, [10]). A language L is called factorizable if any substring of a word $x \in L$ also belongs to L. For a factorizable language L, the set of forbidden words, *i.e.*, the complementary set $L' = \Sigma^* - L$, acquires a minimal property; namely, there is a subset $L'' \subset L'$ of minimal forbidden words that cannot be further cut into shorter words without producing a word in L. A factorizable language L is entirely determined by the subset of minimal forbidden words L'' [10].

Now, we define the language L as the collection of all sequences that have a unique reconstruction at a given K. In other words, L is the set of all uniquely reconstructable sequences. The langauge L is factorizable by construction as any substring of a word $x \in L$ must be uniquely reconstructable; otherwise, x cannot belong to L.

In the case of natural proteins, one always deals with sequences of finite length. Then, the corresponding language L of uniquely reconstructable sequences must be a regular language. Hence a FSA to recognize L exists. In fact, such FSAs have been explicitly built and implemented as computer programs [11]. We note that the above discussion on factorizable langauge applies to any finite alphabet with more than two letters.

V. RESULT OF DATABASE INSPECTION

Equipped with these programs, we can look at real protein data. We studied a subset PDB.SEQ of the SWISS-PROT database [7]. This is a collection of all proteins that have structural data in the PDB database [12]. In the 2005 March Release of PDB.SEQ, there were 6790 protein sequences that did not contain the letter X for undetermined amino acid. In Table 1, we list the number of proteins that satisfy the $R(K) = 1$ relation for the first time in order of increasing $K \leq 30$, the cumulative number of uniquely reconstructable proteins up to the given K, and the cumulative percentage of these sequences among the total of 6790 proteins. It is clearly seen that there is a sharp transition in the percentage of uniquely reconstructable proteins around $K \sim 5$ to 6. At $K = 7$, among the 6790 proteins, 96.28 % are uniquely reconstructable.

Table 1. Number of uniquely reconstructible proteins from a set of 6790 natural proteins.

K	#(proteins) with $R(K) = 1$	Cumulative #(proteins)	Cumulative percentage
2	10	10	0.15
3	82	92	1.35
4	1346	1438	21.12
5	3584	5022	73.78
6	1300	6322	92.88
7	232	6554	96.28
8	82	6636	97.49
9	44	6680	98.13
10	25	6705	98.50
11	10	6715	98.65
12	12	6727	98.82
13	9	6736	98.96
14	7	6743	99.06
15	7	6750	99.16
16	3	6753	99.21
17	4	6757	99.27
18	2	6759	99.29
19	1	6760	99.31
20	5	6765	99.38
21	4	6769	99.44
22	0	6769	99.44
23	2	6771	99.47
24	3	6774	99.51
25	0	6774	99.51
26	0	6774	99.51
27	1	6775	99.53
28	4	6779	99.59
29	1	6780	99.60
30	0	6780	99.60

The non-uniqueness of reconstruction is caused by the presence of scattered repeated peptides in the protein sequence. So far, we have used exact matching of letters to determine repeated nodes. If one is to get closer to biological reality, it is appropriate to introduce fur-

Table 2. Proteins with a big number of reconstructions.

Protein	Number of Amino Acids	K for $R(K) = 1$
APOA_HUMAN	4548	35
RPB1_YEAST	1733	36
ICEN_PSESY	1200	46
NEBU_HUMAN	6669	59
FIB1_PETMA	966	61
MAGA_XENLA	303	84
SRTX_ATREN	543	101
CR1_HUMAN	2039	103
CNA_STAAU	1183	128
CIPA_CLOTM	1853	179

ther coarse-graining by combining amino acids with similar physico-chemical properties. For example, by reducing the 20-letter alphabet to that of 16, combining R with K, F with Y, D with E, and V with I, the number of reconstructions would greatly increase. These inexact matchings, in fact, reflect more biologically meaningful repeats in homologous proteins. By dropping a few sequences with too large a number of nodes (> 200) or with the letter X, we studied 221,415 proteins in the SWISS-PROT database. It turns out that 46.8 % proteins still have a unique reconstruction at $K = 5$.

VI. PROTEINS WITH LARGE NUMBERS OF RECONSTRUCTIONS

It is curious to note that there exists a small set of proteins that have a large number of reconstructions at moderate K and that the K value that makes $R(K) = 1$ may be much greater than 5. Among the 6790 sequences studied in the preceding section, there are 10 such proteins. We list them in Table 2.

Take, for example, the protein ICEN_PSESY in Table 2. This ice nucleation protein of 1200 amino acids has a reconstruction number as huge as $R(11) = 1.56 \times 10^{27}$ at $K = 11$, which is caused by the large number of octopeptide periodicities. Another protein SRTX_ATREN has $R(11) = 9.97 \times 10^5$ caused mainly by 12 tandem repeats, each having 40 almost identical amino acids. These repeats account for the majority of the sequence, *i.e.*, 480 amino acids from a total of 543. Thus, we can pick up some peculiar proteins without any prerequisite biological knowledge by looking for sequences with a large number of reconstructions or by looking at large K values that make $R(K) = 1$.

Correlations in DNA sequences have been widely studied; see, *e.g.*, the review by W. Li [13]. However, unlike DNA sequences, protein sequences are too short to allow for standard correlation studies because it is difficult to define correlation functions by averaging over a long enough sequence. Nonetheless, repeated segments may be considered as the strongest form of correlation. As we pointed out in the preceding sections, the nonuniqueness of sequence reconstructions is caused by scattered repeated peptides. We say “scattered” because simple local repeats will not increase the number of reconstructions. Only the presence of similar, but not identical, peptides at different parts of a protein leads to a combinatorial increase of the number of reconstructions. In this way, repeated peptides in proteins may be classified into simple and complex ones. Our programs may help in picking out proteins with complex repeat structures. Studies along this line are under way.

We note that the first part of this paper has been deposited to arXive.org [14], and Ref. [15] may be considered as a broad introduction to the subject of factorizable language.

ACKNOWLEDGMENTS

We thank Ji Qi and Lei Gao for collaboration on prokaryotic phylogeny, Qiang Li for discussions on the explicit construction of the corresponding FSAs, Chan Zhou and Li Xia for discussion on the sharp transitions in K to unique reconstructions in natural, as well as random amino-acid sequences. The work was partially supported by the State Key Basic Research Project, grant No. TG2000077308, and a grant from Fudan University.

REFERENCES

[1] Ji Qi, Bin Wang and Bailin Hao, J. Mol. Evol. **58**, 1 (2004).

[2] Ji Qi, Hong Luo and Bailin Hao, Nucl. Acids Res. **32** (July 2004), Web Server Issue, W45.

[3] Lei Gao, Ji Qi and Bailin Hao, *AAPPS Bulletin* (June 2006), p. 3.

[4] Fleischner, *Eulerian Graphs and Related Topics*, Part 1, Vol. 2 (Elsevier, 1991), p. IX80.

[5] B. Bollobás, *Modern Graph Theory* (Springer-Verlag, New York, 1998).

[6] J. P. Hutchinson, Utilitas Mathematica **7**, 241 (1975).

[7] The SWISS-PROT database: http://www.expasy.org/sprot/

[8] L. Kontorovich, Theor. Computer Sci. **329**, 271 (2004).

[9] Bailin Hao, Huimin Xie, Zuguo Yu and Guoyi Chen, *Physica* A **288**, 10 (2000).

[10] Huimin Xie, *Grammatical Complexity and One-Dimensional Dynamical Systems* (World Scientific, Singapore, 1996).

[11] Qiang Li, Huimin Xie, *Finite Automata for Testing Uniqueness of Eulerian Trails*, arXive: cs.CC/0507052 (July, 2005).

[12] The RCSB Protein Data Bank: http://www.rcsb.org/pdb/

[13] W. Li, Computer & Chemistry **21**, 257 (1997).

[14] Bailin Hao, Huimin Xie and Shuyu Zhang, *Compositional Representation of Protein Sequences and the Number of Eulerian Loops*, arXive: physics/0103028.

[15] Bailin Hao and Huimin Xie, Int. J. Mod. Phys. B (submitted).