Two Component Systems Sequence Characteristics Identification in Bacterial Genome

Yaw-Ling Lin*

Department of Computer Science and Information Management, Providence University,
200 Chung Chi Road, Shalu, Taichung County, Taiwan 433, R.O.C.
e-mail: yllin@pu.edu.tw

Abstract

2CS (Two component systems) are important molecules for bacteria to detect one or more environmental stimuli and activate the expression of genes necessary for the appropriate response. It is a good research direction to elucidate the functions of bacterial genomes by exploring the functions of 2CS. However, we have little knowledge about 2CS and there is no specific software to discover and predict the virulent genes in the whole genomes.

In this paper, we exploit the algorithms for signature identification in bacterial genomes to realize the functions of virulent genes. Our results can be used to construct the database for bacterial virulence associated genes, which is useful for novel drug designs, for diagnosis of bacterial diseases, and for vaccine development. We will develop a efficient program for searching signatures of bacterial genomes and construct a signature database for bacterial genomes. An efficient program for searching signatures of 2CS will be developed and a signature database for 2CS will be constructed. The results can be used to identify novel 2CS, classify 2CS, and exploit the evolutionary relations among 2CS.

Keywords: two-component systems, bacterial genome, algorithms, sequence characteristics.

1 Introduction

Bacterial infections remain one of the leading causes of morbidity and mortality of humans in the world. Antibiotic is the standard treatment but drug-resistant bacterial strains are common that significantly limited the effectiveness of antibiotics. It is not surprising to find that many companies have exerted tremendous effort to develop novel antibiotics. So far, less than 20 classes of antibiotics and target molecules in bacteria are known. To overcome the drug resistant problem, many underutilized drug targets are being reevaluated and novel targets are in urgent demands.Bacterial components that serve as a target of drug intervention can be divided arbitrarily into several categories. These include virulence factors, gene products essential for the growth during infection, enzymes unique in bacteria, bacterial membrane transporters, bacterial two-component signal transduction pathway, product of genes unique in virulent strains of the bacterial pathogen, and product of genes conserved through evolution. Among these potential drug targets, the bacterial virulence factors are the most obvious and likely to be the most effective targets for antibacterial drug intervention.

Despite approximately 150 antibiotics have been approved by FDA, only few drugs are available for a certain type of infection. In certain particularly resistant bacterial strain, such as methicillin-resistant Staphylococcus aureus (MRSA), only one or two antibiotics (eg. vancomycin and the recently approved oxazolidinon in MRSA) remain effective. In addition, many antibiotics are too toxic or unstable for internal uses that leave a lot of rooms for improvement in their pharmacokinetic and toxicological properties. Although new lines of antibiotic are in urgent demands, there were only 27 antibiotics under development in 1998, most of them focus on modification of existing drugs. It is expected that only a handful of them will be approved by FDA in the near future. Diagnosis of bacterial disease is a rather time-consuming process even in modern clinical laboratories. Typically 3-4 days are required to make a diagnosis for acute bacterial infections and up to 4 weeks for a chronic infection such as tuberculosis. It can take an even longer time if drug susceptibility tests are included. Therefore, it is not unusual that

*The work is supported in part by the National Science Council, Taiwan, R.O.C, grant NSC-90-2213-E-126-013.

physicians could choose the wrong type of antibiotics, which not only delay the timing of appropriate treatment but also can result in drug resistance. How to differentiate critical groups of pathogen and together their drug susceptibility pattern within hours has become an important research direction in clinical microbiology.

Among numerous bacterial species exist on earth, only less than a hundred can cause human diseases and no more than 50 are commonly encountered in clinical laboratories. It becomes evident that these pathogens do not cause diseases accidentally but have been gone through a rather long evolution process. All pathogens evolved a battery of virulence genes during interacting with hosts. In addition, more virulence-associated genes were recruited from many other bacterial species. The accumulation of a large number of virulence genes significantly increases the capability of the pathogen to adapt and propagate in the hosts.

These virulence genes are therefore the prime targets for development of diagnostic tool and vaccine, and for antimicrobial drug intervention. In tradition, the bacterial virulence factors were identified through a series of microbiology and immunology studies. This process has been greatly facilitated recently by the completion of genome projects of many pathogenic bacteria. Many virulence-associated genes can be readily identified through bioinformatic approaches. Nevertheless, the BLAST programs [1] commonly used in genome analysis have their limitation. On the basis of BLAST search, it is estimated that approximately 20% of genes found in genome programs are novel sequences. Therefore, how to develop a novel annotation tool to identify the possible functional roles of these genes becomes a very important task. In addition, since the functions of these novel sequences are yet to be identified, very little attention has been drawn on these sequences. Therefore, these sequences are suitable for new players in bioinformatic area such as research groups in Taiwan to explore. All bacterial genome databases, including those of H. pylori, can be accessed or downloaded through The Institute of Genome Research (http://www.tigr.org) or National Center for Biotechnology Information, USA (http://www3.ncbi.nlm.nih.gov). The genome of K. pneumonia MGH 78578 has been completed by the Genomic Sequencing Center at Washington University, St. Louis. The nucleotide sequence of contigs larger than 1 kb in size can be obtained from the web site http://genome.wustl.edu/gsc/bacterial/klebsiella/klebsiella.shtml and will soon be available in GenBank/EMBL database.

Rapid adaptation to environmental challenge is essential for bacterial survival. To orchestrate their adaptive responses to changes in their surroundings, bacteria mainly use so-called 'two-component regulatory systems' (2CS) [3]. These systems are usually composed of a sensor kinase, which is able to detect one or several environmental stimuli, and a response regulator, which is phosphorylated by the sensor kinase and which, in turn, activates the expression of genes necessary for the appropriate physiological response. Sensor kinases (or histidine kinases) usually possess two domains: an input domain, which monitors environmental stimuli, and a transmitter domain, which auto-phosphorylates following stimulus detection. The input domain varies in length and amino acid sequence from one histidine kinase to another, conferring specificity for different stimuli. By contrast, the transmitter domain shows high sequence conservation. It contains an invariant histidine residue that is phosphorylated in an ATP-dependent manner and short stretches of conserved amino acids, in particular two glycine-rich motifs involved in ATP binding (the NG1FG2 motif). A classical response regulator contains an amino-terminally located conserved receiver domain that is phosphorylated by the sensor kinase at a strictly conserved aspartate residue, leading to activation of the carboxy-terminal effector or output domain [5, 6].

Because of its important functional roles and ubiquitous nature in most bacterial and fungal species, 2CS have been considered as very good targets in drug development [2]. In addition, 2CS also meet the following criteria for drug development. Some of 2CS are critical for bacterial growth and coordinate pathogenesis, including some problematic infection (eg. Biofilm formation). The enzymatic activity of 2CS is assayable and homology is high at active site, which lend itself to drug screening. 2CS are not found in humans that provides selective basis over mammalian targets/processes. They are surface exposed and are previously unexploited targets. Finally, there are multiple sets of 2CS in a bacterial genome and hence with low expected resistance. Analysis of complete bacterial genome sequences have shown that the number of these systems varies considerably from one species to the next, from 0 in Mycoplasma spp., 38 in the cyanobacterium Synechocystis [4], and 63 in Pseudomonas aeruginosa [7]. The functional role of most of the 2CS, however, remains elusive. For examples, among the 63 2CS in P. aeruginosa, only 10 or so have been characterized [6].

The identification of the function of these 2CS

would greatly facilitate not only our understanding on the basic physiology and regulatory networks of bacteria but also designing a way to prevent from causing disease in humans.

2 Co-evolution analysis of sensor and regulator gene in a 2CS

Traditionally transcription regulators are classified according to their helix-turn-helix DNA binding motif and are assigned into families such as LysR or LuxR. Most of the genes encoding the transcription regulator are located in the upstream of their target genes and are transcribed from a divergent promoter in a direction opposite from that of targeted genes. Sequence analysis of the transcription regulators indicates that they are most likely derived from duplication events from an ancestor and was later recruited and clustered together with the target genes. Therefore, in this type of gene cluster, the transcription regulator genes were evolved independently from their target genes. In contrast, the target genes regulated by transcription (response) regulator of a 2CS are generally scattered in the genome, whereas the gene encoding response regulator and the sensor are located within an operon.

It is therefore interesting to know whether the gene encoding regulatory protein and the gene encoding the sensor kinase in a 2CS were derived by duplication from an existing 2CS (the *co-evolution*) or they were evolved independently and were assembled by recombination event later.

To address this question, we will need to know the evolutionary distance of each regulatory protein encoding gene and each sensor-encoding gene of the 63 2CS. The two trees will then be integrated and those sensor and regulator genes exhibit distinct relationship in a 2CS will be selected and analyzed further. In this case, integration of the two trees not only is the key to reveal the secret of 2CS evolution, but also a challenging question in computational biology. One way of extracting the useful clustering information that might later lead to functional classifications of these 2CS from the regulator tree and the sensor tree is to incorporate the evolutionary information from both trees. We consider the following combinatorial problems:

Definition 1 (k-agreement) *Given n terminal nodes that represents n abstract objects (in this case, they are 2CS sequences), these n nodes constitute the same set of leaf nodes within two given topologically different evolutionary/ phylogenic trees, say T_1 and T_2. We call these two trees as* dual trees. *Note that the deletion of any edge separates the tree into two disconnected subtrees, whose leaf nodes are exactly two subsets of the leaf nodes, forming two partitions. The problem is, for a given parameter k, to identify whether there is an edge e_1 in T_1 and e_2 in T_2 whose deletion form the same partition of the leaf nodes such that one of the two partitions has exact size of k.*

ALGORITHM **AGR**(T_1, T_2, n, k)

Input: A sensor kinase tree T_1, a (response) regulator tree T_2, with n leaves.

Output: A list of k pairs of co-subtree (t_1, t_2)'s where t_1 (t_2) is a subtree of T_1 (T_2). These k co-subtrees possess the smallest mutual exclusive ratio.

Step 1: Let the $A = \{a_1, a_2, \ldots, a_{n-1}\}$ denote the $n-1$ subtrees of T_1 defined by the $n-1$ internal nodes of T_1. Let $B = \{b_1, b_2, \ldots, b_{n-1}\}$ denote the $n-1$ subtrees of T_2. Let output list $P \leftarrow \emptyset$.

Step 2: For each $(a, b) \in A \times B$, compute the mutual exclusive ratio $\Delta(a, b) = |L(a) \cup L(b) \setminus L(a) \cap L(b)| / (|L(a)| + |L(b)|)$.

Step 3: Select the k co-subtree with smallest mutual exclusive ratio among all $\Delta(a, b)$'s. Output these k pairs in non-decreasing order.

END OF **AGR**

Figure 1: Computing the mutual exclusive ratio $\Delta(t_1, t_2)$'s within a pair of co-evolution trees.

Note that it is trivial to identify 1-agreement (as well as n-agreement) nodes from any given dual trees. On the other hand, not every parameter k leads to a feasible solution. The rationale behinds the problem is the following.

Assuming that the dual trees do process a k-agreement subset (with size k) of 2CS sequences, it follows that these k sequences have a very good chance of forming a reasonable candidate for the clustering group, and hopefully these genes would be functionally related to each others. Computationally, the k-agreement problem can trivially solved by a polynomial time algorithm.

There are exactly $(n-1)$ internal nodes for a n-leaf (binary) tree; thus the possible ways of deletion are $n-1$. It follows that arbitrarily picking two pairs of internal nodes, each from one of dual trees, constitutes totally $O(n^2)$ possible pairings. For each

pairing, a linear time algorithm can be used to double check whether these two pairs form an agreement. The total time needed for the trivial algorithm will be $O(n^3)$. More efficient algorithms are attainable for this problem.

Besides the exact match solutions for the k-agreement problem, it will be useful for biologists to have some sort of *fuzzy* measurement of two clusters of genes. Let $A, B \subset S = \{1, 2, \ldots, n\}$ be two clusters of genes set S. The *mutual exclusive ratio*, $\Delta(A, B)$, is defined as the following:

$$\Delta(A, B) = \frac{|A \cup B \setminus A \cap B|}{|A| + |B|}$$

The mutual exclusive ratio between two sets is considered to be a rough measurement of the *degree of difference* between them. Note that $0 \leq \Delta(A, B) \leq 1$; $\Delta(A, B) = 0$ if $A = B$, and $\Delta(A, B) = 1$ if $A \cap B = \emptyset$. In other words, the smaller value of $\Delta(A, B)$ implies a greater *similarity* of A, B. Our algorithm of computing all the mutual exclusive ratios of pairs of subtrees of regulation and sensor trees is illustrated at Figure 1.

3 Correlation analysis of sensor and regulator genes

Gene duplication event is commonly occurred in bacteria that generate many gene families. These gene duplication events raise a very interesting question: does gene duplication tend to occur within a relative short distance on a bacterial genome? Despite this is a reasonable assumption; there has not been solid evidence to support this notion presumably due to lack of suitable testing systems. In bacteria, the number of member in a bacterial gene family is low and the repetitive sequences are too conserved, both are non-informative in genetic analysis. With more than 60 different 2CS in P. aeruginosa genome, it provides a great opportunity for us to test this interesting hypothesis. In this study, a dot-matrix plot will be created, with the X-axis being the physical distance, and Y-axis being the evolutionary distance, between two comparing 2CS.

It is possible that, instead of all 2CS whose sequences possess a correlation between their physical and evolutionary distances. Some subset of 2CS, presumably functionally related, could possess the correlation between their physical and evolutionary distances. Identifying these measurement-correlated groups could be a computational consumption problem. The following combinatorial optimization problem is considered:

Definition 2 (k-correlation) *Consider n nodes being associated with two different distance measurements M_1 and M_2; an $n \times n$ squared distance matrix represents each measurement. These two measurements are called* dual measures. *The difference of two squared matrix can be defined by $d(M_1, M_2) = |M_1 - M_2|$, or some other bio-meaningful functions. Note that a selection of k nodes, $A \subset S$, from the n-set S, produces an induced measurement from the given distance matrix; the induced matrix is denoted by M^A. The problem is, for a given parameter k, to identify the k-node, $A \subset S$, such that the difference $d(M_1^A, M_2^A)$ is minimized.*

In our study, we have shown that this combinatorial problem is intractable even when the distances measurements only consist of two different real numbers. However, it will be interesting to know whether the problem can be reasonably approximated under some special constraints. Note that our k-agreement problem can be a good candidate of approximating the k-correlation problem if we first construct two phylogenic trees from the two given measurements and then apply the k-agreement algorithm for finding probable k's.

4 Concluding Remarks

In this paper, we present results in applying information technologies to construct a diagnostic tool for bacterial diseases and a bacterial virulence associated genes database that can be used as a basis for development of antimicrobial drugs and vaccines. Identification of signatures in bacterial genomes. The signatures for bacterial genomes will be used to build a diagnostic tool for identifying a designated pathogenic bacterium.

Several interesting topics to be discussed in the future include identifying novel 2CS in other bacteria genomes as well as in eucaryotic genomes, clustering analysis of 2CS for functional prediction of uncharacterized genes, and 2CS co-evolutionary analysis of 2CS.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tools. *J. Mol. Biol.*, 215:403–410, 1990.

[2] J. F. Barret and J. A. Hoch. Two-component signal transduction as a target for microbial anti-

infective therapy. *Antimicrob. Agent. Chemo.*, 42:1529–36, 1998.

[3] J.A. Hoch and T.J. Silhavy. *Two-Component Signal Transduction*. ASM Press, 1995.

[4] T. et al. Mizuno. Compilation of all genes encoding bacterial two-component transducers in the genome of the cyanobacterium, synechocystis sp. strain PCC 6803. *DNA Res.*, 3:407–414, 1996.

[5] J.S. Parkinson and E.C. Kofoid. Communication modules in bacterial signalling proteins. *Annu. Rev. Genet.*, 26:71–112, 1992.

[6] A. Rodrigue, Y. Quentin, A. Lazdunski, V. Méjean, and M. Foglino. Two-component systems in pseudomonas aeruginosa: why so many? *Trends Microbiol.*, 8:498–504, 2000.

[7] C.K. et al. Stover. Complete genome sequence of pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, 406:959–964, 2000.