The Compositive Analysis of Two-component Regulatory Systems in Bacterial Genome

Tze-Wei Huang [*] Po-Shun Yu [†] Hsun-Chang Chang [‡] Yaw-Ling Lin [§]

Abstract

Most bacteria adapt to their surroundings by means of ”two-component regulatory systems”. Bacterial infections remain one of the leading causes of morbidity and mortality of humans in the world. The investigation of two-component regulatory systems would greatly contribute to the understanding of bacterial physiology and pathogenesis. A two-component regulatory system is composed of several functional domains. Profile hidden Markov models are probably the most popular application of hidden Markov models in molecular biology at the moment. By applying profile HMMs technology, we propose methods to report and recognize the positions and types of two-component regulatory systems in bacterial genomes by constructing and integrating similarity scores and Bayesian posterior probabilities of functional domains.

Keywords: Two-component regulatory system, composite analysis, functional domains, profile hidden Markov models.

1 Introduction

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” [7]. 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 [18].

In this paper, we propose a new method to recognize positions and types of two-component regulatory systems in a bacterial genome by constructing and integrating similarity score diagrams of 2CS’s functional domains. In the process of constructing these diagrams we (1) collect and classify the 2CS gene sequences according to different subfamilies respectively based on the functional and structural knowledge, (2) MSAs (multiple sequence alignment) are made respectively for sequences of each of the subfamilies, (3) sequences corresponding to each functional domain are trimmed from the MSA according to the known domain information, and (4) build a profile HMM database from trimmed aligned sequences of 2CS’s functional domains. We design a window sliding method scanning from the beginning to the end of a bacterial genome. Along the way the window sliding, we calculate a similarity score (bit score or Bayesian posterior probability) for the window of being a specific type of 2CS’s functional domains in that position of the genome. Then we can construct a bit score diagrams for every 2CS’s functional domain along the different positions on the query genome. In terms of integrating these diagrams we can recognize the positions and types of two-component regulatory systems in a bacterial genome by observing and analyzing the interactions between these curves from these diagrams.

To proceed with our experiments, we take the genomic sequence of *Pseudomonas aeruginosa* as the target genome. *Pseudomonas aeruginosa* is well studied, and it is one of the most important human pathogen as it is commonly responsible for respiratory-tract infections in cystic fibrosis patients as well as nosocomial infections in immunocompromised patients following surgery or

[*] Department of Comput. Sci. and Info. Management, Providence University, e-mail: g9134012@cs.pu.edu.tw

[†] Department of Comput. Sci. and Info. Management, Providence University, e-mail: peteryu@cs.pu.edu.tw

[‡] Department of Comput. Sci. and Info. Management, Providence University, e-mail: hcchang@cs.pu.edu.tw

[§] Department of Comput. Sci. and Info. Management, Providence University, e-mail: yllin@pu.edu.tw (The work is supported in part by the National Science Council, Taiwan, R.O.C, grant NSC 92-2213-E-126-001.)

the administration of cytotoxic drugs, or patients with burn wounds [14]. *P. aeruginosa* produces a wide variety of exoproducts, many of which contribute to the virulence of this opportunistic pathogen [18]. The investigation of 2CS (two-component regulatory system) will greatly contribute to the understanding of bacterial physiology and pathogenesis. The complete genome sequence of *P. aeruginosa* strain PAO1 has been released [21].

2 Method

Hidden Markov models (HMM) have been applied successfully over the last two decades in a wide variety of speech and language related recognition problems including speech recognition [10], named entity finding [1], optical character recognition [11], and topic identification [19]. A hidden Markov model is defined by a set of output symbols, a set of states, a set of probabilities for transitions between the states, and a probability distribution on output symbols for each state [16].

Functional biological sequences typically come in families, and many of the most powerful sequence analysis methods are based on identifying the relationship of an individual sequence to a sequence family. If you already have a set of sequences belonging to a family, you can perform a database search for more members using pairwise alignment with one of the known family members as the query sequence. To be more thorough, you could even search with all the known members one by one. However, pairwise searching with any one of the members may not find sequences distantly related to the ones you have already. An alternative approach is to use statistical features of the whole set of sequences in the search. Similarly, even when family membership is clear, accurate alignment can be often be improved significantly be concentrating on features that are conserved in the whole family. We know that a multiple alignment can show how the sequences in a family relate to each other. Here we use a particular type of hidden Markov model called profile HMM which is well suited to modeling multiple alignments. Profile HMMs are probably the most popular application of hidden Markov models in molecular biology at the moment [3].

Here we present the flow diagram of our method as illustrated at Figure 1.

2.1 Classifying the 2CS Gene Sequences

We first collect an inventory of 2CS gene sequences of *Pseudomonas aeruginosa.* This goal can be achieved simply by browsing the web pages for the gene sequences classified by functions in the *Pseudomonas* genome project database [21] and then downloading the sequences from the category of "two-component regulatory systems". With the typical 2CS gene sequences in hands, we can classify the 2CS gene sequences according to different subfamilies respectively based on the functional and structural knowledge according to the annotations [20] to these gene sequences.

We apply Dr. Rodrigue's research [18] that there are putative genes encoding 63 sensor kinases and 64 response regulators. The Figure 2 shows the number of genes belonging to a specific family, and the Figure 3 shows the subfamilies of sensor kinases. In Figure 4 we lists the subfamilies of response regulators. The following subsections describe subfamilies of two-component systems in *Pseudomonas aeruginosa* in details.

2.1.1 Sensory kinases

Classical sensory kinases are composed of a transmitter domain preceded by an amino-terminal input domain. Sequence analysis of the 63 potential histidine kinases identified in the genome of *P. aeruginosa* revealed 42 putative classical sensory kinases. The *P. aeruginosa* genome also contains five histidine kinase genes that are homologous to the chemotaxis protein CheA. These kinases differ from other histidine kinases as the transmitter domain is located near the amino terminus and the amino acid sequence surrounding the conserved histidine is more closely related to histidine containing phosphotransfer module (Hpt) domains than to classical transmitters [2]. More complex histidine kinases also possess a receiver domain adjacent to the transmitter domain. This receiver domain is similar to those of response regulators and is linked to an Hpt module. We have identified five *P. aeruginosa* genes that could encode unorthodox histidine kinases (ORFs 42081, 40722, 41506 and 41910, and gacS) in which the three phosphotransfer domains (i.e.transmitter, receiver and Hpt) are combined. The five Hpt-domain sequences can be aligned with the corresponding domain of known unorthodox kinases such as TorS and

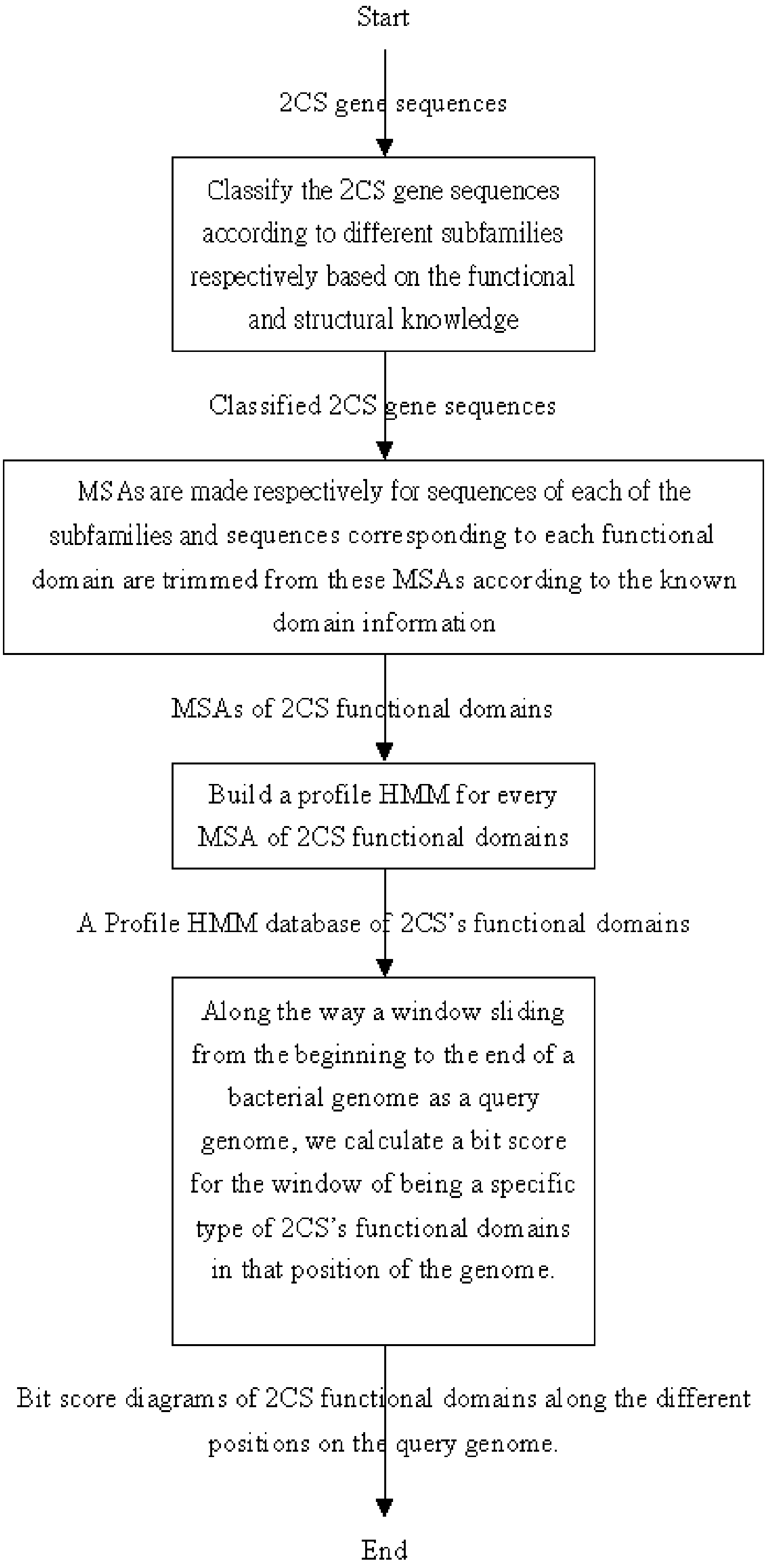
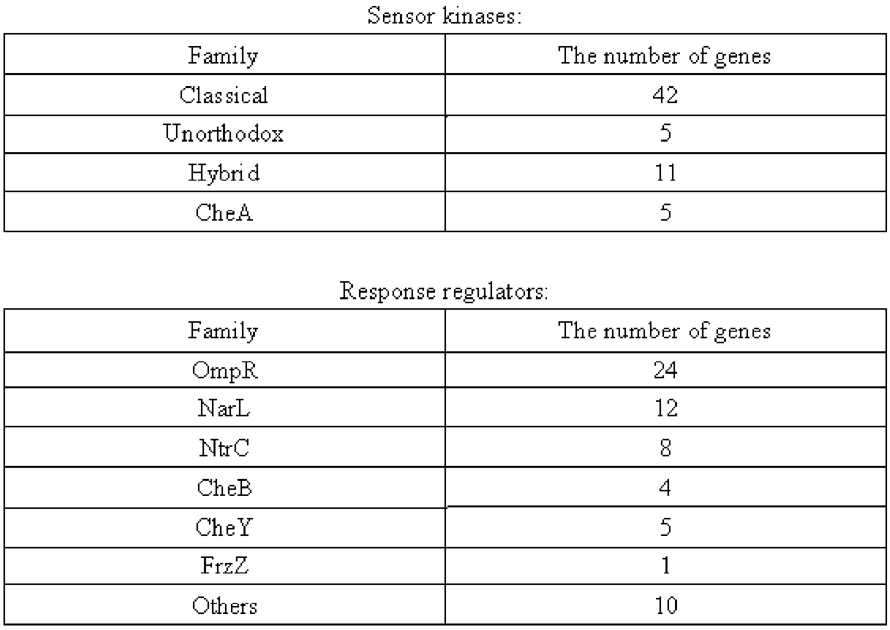

Start
2CS gene sequences
Classify the 2CS gene sequences according to different subfamilies respectively based on the functional and structural knowledge
Classified 2CS gene sequences
MSAs are made respectively for sequences of each of the subfamilies and sequences corresponding to each functional domain are trimmed from these MSAs according to the known domain information
MSAs of 2CS functional domains
Build a profile HMM for every MSA of 2CS functional domains
A Profile HMM database of 2CS's functional domains
Along the way a window sliding from the beginning to the end of a bacterial genome as a query genome, we calculate a bit score for the window of being a specific type of 2CS's functional domains in that position of the genome.
Bit score diagrams of 2CS functional domains along the different positions on the query genome.
End

Figure 1: The flow diagram of our method.



Sensor kinases:

Family	The number of genes
Classical	42
Unorthodox	5
Hybrid	11
CheA	5

Response regulators:

Family	The number of genes
OmpR	24
NarL	12
NtrC	8
CheB	4
CheY	5
FrzZ	1
Others	10

Figure 2: The number of genes belonging to a specific family.

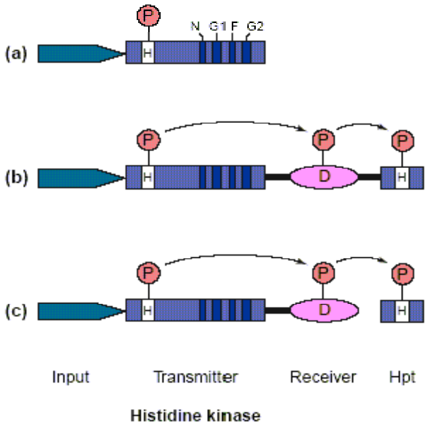
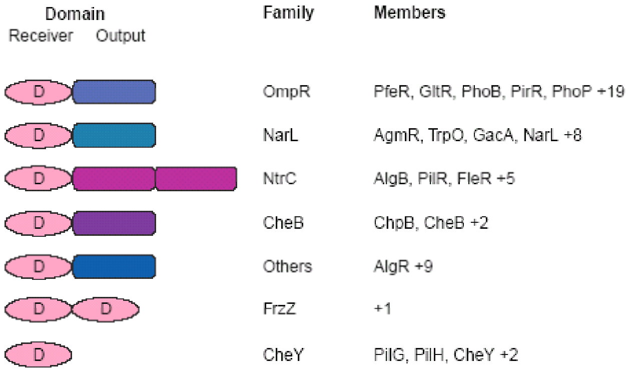


Figure 3: The subfamilies of sensor kinases. (a)classical (b)unorthodox (c)hybrid



Family	Members
OmpR	PfeR, GltR, PhoB, PirR, PhoP +19
NarL	AgmR, TrpO, GacA, NarL +8
NtrC	AlgB, PilR, FleR +5
CheB	ChpB, CheB +2
Others	AlgR +9
FrzZ	+1
CheY	PilG, PilH, CheY +2

Figure 4: The subfamilies of response regulators. D is the strictly conserved aspartate residue, and +n is the number of potential response regulators that were previously unknown.

ArcB of E. coli and BvgS of Bordetella pertussis [9, 8, 24]. The other 11 sensory kinases are receiver-containing hybrid kinases that do not contain an extended carboxy terminus in the receiver domain. In one example, the transmitter domain is followed by two tandem receiver domains (ORF 41634) [17]. The phosphorelay from such hybrid kinases has been studied in detail in *Saccharomyces cerevisiae* [15] and *Vibrio harveyi* [6].

2.1.2 Response regulators

The output domains of response regulators can be classified into different subfamilies on the basis of whether they are similar to the E. coli OmpR, NarL or NtrC output domains [13]. Examination of these domains in *P. aeruginosa* response regulators revealed that 24, 12 and eight are similar to OmpR, NarL and NtrC, respectively. These 44 open reading frames (ORFs) are potentially DNA-binding transcriptional regulators, as these three *E. coli* protein families are DNA-binding proteins. Among the 20 other *P. aeruginosa* response regulators, four can be classified in the CheB subfamily, in which the output domain has esterase activity [12]; five resemble CheY in that they lack an output domain; and one has two receiver domains side by side and thus lacks an output domain, as does FrzZ in *Myxococcus xanthus* [23]. Finally, 10 response regulators possess output domains that share only weak similarities with other known response regulators and with each other.

2.2 Locating the Functional Domains by MSA

We use the ClustalW package [22] to MSA our family sequences. Then we trim these MSAs into smaller ones for each of them as a functional domain MSA. In order to trim these MSAs properly, a lot of literatures concerning 2CS gene structures have been discussed in the past. Here we consider parts of the MSAs from the gene families has been properly trimmed by biologists according their biological functional or structural evidences.

2.3 Build a profile HMM database from MSAs of Functional Domains

Since we have MSAs for different functional domains of 2CS in bacterial. For every MSA from sequences belonging to the same functional domain, we construct a profile HMM using the `hmmbuild` program in the HMMER profile HMM package. Developed chiefly by Sean Eddy, the HMMER package [4] is freely available under the GNU General Public License and includes the necessary model-building and model-scoring programs relevant to homology detection. In addition, the package contains a program that calibrates a model by scoring it against a set of random sequences and fitting an extreme value distribution to the resultant raw scores; the parameters of this distribution are then used to calculate accurate E-values for sequences of interest. All HMMER models used in this study were calibrated in this way [5].

2.4 Prediction and Classifying Positions and Types of Domain Families

Here we design a window sliding from the beginning to the end of a bacterial genome. Assume that the window size is s bps wide and slide forward k bps every time. Along the way the window sliding, we calculate a bit score for the window of being a specific type of 2CS's functional domains in that position of the genome by using `hmmsearch` program in the HMMER package. Then we can construct a similarity score diagrams of 2CS's functional domains along the different positions on the query genome. In terms of integrating these diagrams we can recognize the positions and types of two-component regulatory systems in a bacterial genome by observing and analyzing the intersections between these curves from these diagrams. The following are details of bit score calculation [5]:

The bit score is a log-odds score in log base two (thus, in units of bits). Specifically, it is:

$$s = \log_2 \frac{P(seq|H)}{P(seq|null)}$$

$P(seq|H)$is the probability of the target sequence according to your profile HMM. $P(seq|null)$ is the probability of the target sequence given a "null hypothesis" model of the statistics of random sequence. In HMMER, this null model is a simple one-state HMM that says that random sequences are i.i.d. sequences with a specific residue composition (this "null model distribution" is part of the profile HMM save file, and it can be altered when you do an `hmmbuild`). Thus, a positive score means the profile HMM is a better model of the target sequence than the null

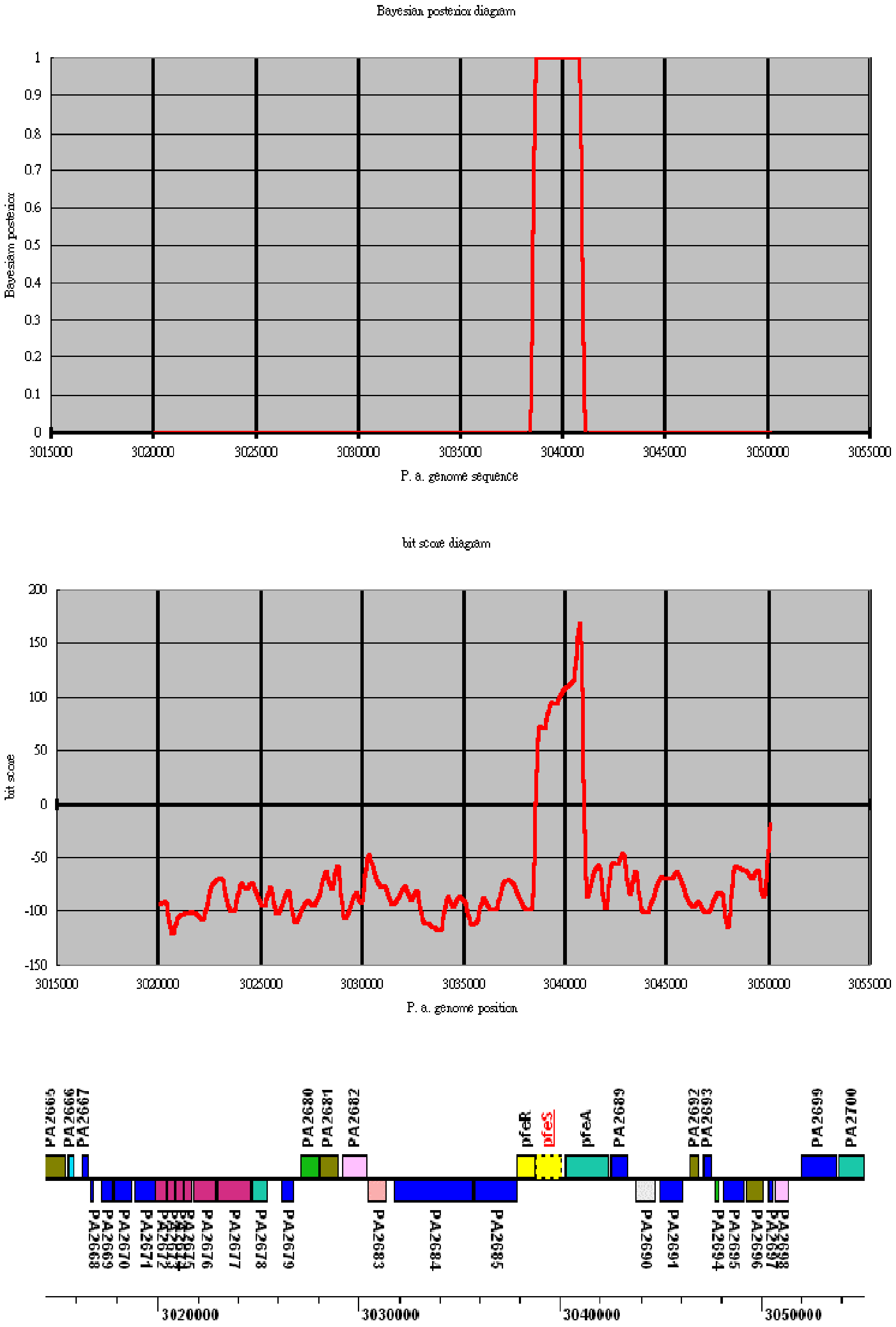


Figure 5: The bit score and Bayesian posterior diagram effectively reflect the position of PA2687 two-component sensor PfeS [21].

model is (e.g. the profile HMM gives a higher probability).

If we take the Bayesian view, we're interested in the probability of a hypothesis H given some observed data *seq*:

$$P(H|seq) = \frac{P(seq|H)P(H)}{\sum_{H_i} P(seq|H_i)P(H_i)},$$

an equation which forces us to state explicit probabilistic models not just for the hypothesis we want to test, but also for the alternative hypotheses we want to test against. Up until now, we've considered two hypotheses for an observed sequence seq: either it came from our profile HMM, or it came from our null hypothesis for random, unrelated sequences (call that model null). If these are the only two models we consider, the Bayesian posterior for the model H is:

$$P(H|seq)$$
$$= \frac{P(seq|H)P(H)}{P(seq|H)P(H) + P(seq|null)P(null)}$$

Recall that the log odds score reported by HMMER's alignment algorithms is

$$s = \log_2 \frac{P(seq|H)}{P(seq|null)},$$

and let's assume for simplicity that a priori, the profile and the null model are equiprobable, so the probabilities $P(H)$ and $P(null)$ cancelled with each other. Then the log -odds score s is related to the Bayesian posterior by the *sigmoid* function,

$$P(H|seq) = \frac{e^s}{e^s + 1}.$$

See [5, 4, 3] for the details and validations of the discussions.

3 Preliminary result

As we mention in Section 2.2, trimming subfamilies MSAs remains a tedious and time-consuming process for the functional biologists. Here we use a MSA of sensor kinase gene sequences in P. A. as a profile HMM training data to construct a sensor kinase gene profile HMM. We use this profile HMM to serve as one of the functional domain profile HMMs that we are interested. This profile HMM is used to validate our computation mechanism correctly operates as expected. The bit score and Bayesian posterior diagram with the window size is 3000 bps wide and slide forward 300 bps every time. Note that Figure 5 does effectively reflect the position of PA2687 two-component sensor PfeS [21]. Totally we calculate 20873 bit scores and Bayesian posteriors along the P. A. genome with length about 6263000 bps.

4 Conclusions

We propose methods for recognizing the positions and types of two-component regulatory systems in a bacterial genome by constructing and integrating similarity score diagrams of 2CS's functional domains. Currently, we are still constructing websites to help biologists to recognize the positions and types of two-component regulatory systems as well as other domains signatures in a bacterial genome. Our system will be designed to have the function of estimating probabilities in which a recognized domains family belong, and try to explain this result according to surveyed 2CS gene structure literature. Most of the materials and some of our preliminary results can be obtained in our website [25].

Acknowledgements

We would like to thank Dr. Yin-Te Tsai for helpful comments and suggestions. We also thank Dr. Hwan-You Chang, Dr.Hwei-Ling Peng, and her PhD student Ying-Tsong Chen for providing meaningful biological knowledge about two-component regulatory systems in bacteria.

References

[1] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: a high-performance learning name finder. *Fifth Conference on Applied Natural Language Processing(published by ACL)*, pages 194–201, 1997.

[2] A.M. Bilwes et al. Structure of chea, a signaltransducing histidine kinase. *Cell*, 96:131–141, 1999.

[3] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1998.

[4] Sean Eddy. Hmmer: profile hmms for protein sequence analysis. `http://hmmer.wustl.edu./`.

[5] Sean Eddy. *HMMER User's Guide version 2.3.1.* Howard Hughes Medical Institute and Dept. of Genetics Washington University School of Medicine, May 2003.

[6] J.A. Freeman and B. Bassler. Sequence and function of luxu: a two-component phosphorelay protein that regulates quorum sensing in vibrio harveyi. *J. Bacteriol.*, 181:899–906, 1999.

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

[8] S. Iuchi et al. The arcb gene of escherichia coli encodes a sensor-regulator protein for anaerobic repression of the arc modulon. *Mol. Microbiol.*, 4:715–727, 1990.

[9] C. Jourlin and otheres. Transphosphorylation of the torr response regulator requires the three phosphorylation sites of the tors unorthodox sensor in escherichia coli. *J. Mol. Biol.*, 267:770–777, 1997.

[10] J. Makhoul and R. Schwartz. State of the art in continuous speech recognition. *Proc. Natl. Acad. Sci. USA 92*, pages 9956–9963, 1995.

[11] J. Makhoul, R. Schwartz, C LaPre, and I. Bazzi. A script-independent methodology for optical character recognition. *Pattern Recognition*, Vol 31, No.9:1285–1294, 1998.

[12] J.S. Parkinson. Signal transduction schemes of bacteria. *Cell*, 73:857–871, 1993.

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

[14] M. Pollack. Pseudomonas aeruginosa. *In Principles and Practices of Infectious Diseases (Mandell, G.L. et al., eds)*, pages 1673–1691, Churchill Livingstone, 1990.

[15] F. Posas et al. Yeast hog1 map kinase cascade is regulated by a multistep phosphorelay mechanism in the sln1-ypd1-ssk1 two-component osmosensor. *Cell*, 86:865–875, 1996.

[16] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE 77*, pages 257–286, 1989.

[17] A. Rodrigue. Presentation of 2cdb: a database of two-component systems in pseudomonas aeruginosa. `http://ir2lcb.cnrs-mrs.fr/Pa2Cdb/`.

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

[19] R. Schwartz, T. Imai, F. Kubala, L. Nguyen, and J. Makhoul. A maximum likelihood model for topic classfication of broadcast news. *Proc. Eurospeech '97, Rhodes, Greece*, pages 1455–1458, 1997.

[20] Stover et al. Pseudomonas aeruginosa community annotation project. `http://www.cmdr.ubc.ca/bobh/PAAP.html`.

[21] Stover et al. Pseudomonas genome project. `http://www.pseudomonas.com/`.

[22] J.D. Thompson, D.G. Higgins, and T.J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.

[23] K.G. Trudeau et al. Identification and characterization of frzz, a novel response regulator necessary for swarming and fruiting-body formation in myxococcus xanthus. *Mol. Microbiol.*, 20:645–655, 1996.

[24] M.A. Uhl and J.F. Miller. Integration of multiple domains in a two-component sensor protein: the bordetella pertussis bvgas phosphorelay. *EMBO J.*, 15:1028–1036, 1996.

[25] Po-Shun Yu, Hsun-Chang Chang, and Tze-Wei Huang. Providence university bioinformatics forum. `http://bioinfo.cs.pu.edu.tw/`.