

Manifestation and Exploitation of Invariants in Bioinformatics

Limsoon Wong

School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543
`wongls@comp.nus.edu.sg`

Abstract. Whenever a programmer writes a loop, or a mathematician does a proof by induction, an invariant is involved. The discovery and understanding of invariants often underlies problem solving in many domains. I discuss in this tutorial powerful invariants in some problems relevant to biology and medicine. In the process, we learn several major paradigms (invariants, emerging patterns, guilt by association), some important applications (active sites, key mutations, origin of species, protein functions, disease diagnosis), some interesting technologies (sequence comparison, multiple alignment, machine learning, signal processing, microarrays), and the economics of bioinformatics.

1 Introduction

The frontier of biological and medical sciences is exciting and full of opportunities today, due to the accumulation of huge amount of biomedical data and the imminent need to turn such data into useful knowledge [31]. There are numerous techniques for dealing with each of the broad spectrum of bioinformatics problems that have emerged, and more are being proposed everyday. There have been a number of useful reviews and tutorials written on various bioinformatics problems. In general, these reviews and tutorials are focused on a specific bioinformatics problem [5], or on a specific technology [19], or both [16].

In this tutorial, I do not focus on a single problem or a single technology. Instead, I present a large varieties of problems and techniques, and try to highlight a fundamental property that is common to all of them. Specifically, I observe that these problems are characterized by *invariants* that emerge naturally from the causes and/or effects of these problems, and show that the techniques for their solutions are essentially exploitation of these invariants.

Before I provide more detail, let me first use an example to illustrate the concept of invariants. We are given a bag of x red beans and y green beans. We are to repeatedly remove two beans from the bag. If both beans are red, we discard both of them. If both beans are green, we discard one and return the other one to the bag. If one is green and one is red, we discard the green bean and return the red bean to the bag. Suppose there is a single bean left in the bag at the end of this process. Can we predict the color of this last remaining

bean? The solution is simple: This last remaining bean is red if and only if x is odd. The simplicity of this solution arises from a property of the process: The parity of the red beans is preserved—i.e., invariant—at each step of the process. We thus see that invariants are fundamental properties of a problem and can be exploited to provide surprisingly simple solutions to the problem.

As mentioned earlier, the problems presented in this tutorial are all manifestations of invariants. Specifically,

- Section 2 and Section 3 look at the problems of recognizing the active sites of an enzyme, finding the mutations that reduces the efficiency of a protein function, and determining the origin of Polynesians. These problems are manifestations of invariants in the process of Evolution—in particular, sequence features that are conserved during evolution.
- Section 4 looks at the problem of protein function prediction. The process of Evolution has also preserved and/or imposed a number of invariant characteristics on proteins with different functions. The invariant characteristics of a protein is naturally useful for prediction of its function.
- Section 5 looks at the problem of disease subtype diagnosis. Each disease and its various subtypes have their underlying causes. The causes are often difficult to decipher due to the complexity of molecular circuitries and gene-environment interactions. Nevertheless, different causes have different invariant down-stream effects that are useful as diagnostic indicators.

I also show that the techniques for their solutions are essentially exploitation of these invariants.

2 Invariants in Evolution

Let me begin with the problem of finding active sites of an enzyme. An “active site” is a region of an enzyme that a substrate binds to, so that a biochemical reaction can occur. Such sites must be conserved through the evolution process, because the function of the enzyme would be disabled, severely reduced, or completely changed if the physico-chemical properties of the amino acid sequence at these sites were changed. That is, the physico-chemical properties of the amino acid sequence required at these sites are the invariants of the enzyme that must be preserved during the evolution process in order for the protein to retain its specific enzyme function.

Figure 1 illustrates the evolution of a hypothetical enzyme. The function f of ancestor enzyme #1 is characterized by active site “A”. Enzyme #2 is evolved from enzyme #1 by having a different physico-chemical property “ a_1 ” at the site “A”; thus it no longer has function f . Enzyme #3 is also evolved from enzyme #1, but by having a different physico-chemical property at site “B”; thus it may have a new function g in addition to f . Enzymes #4, #5, #6, and #7 are similarly evolved. It is clear that “A” is the only property common between all enzymes that have function f . Similarly, “A” and “B” are the only properties common between all proteins that have both functions f and g .

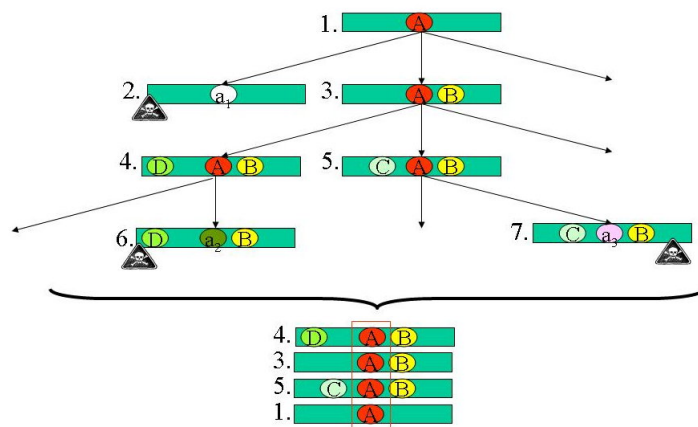


Fig. 1. The evolution of a hypothetical enzyme.

The effect of this type of requirement is that the amino acid sequence at these sites is also under pressure to be invariant. This invariant is an indirect and imperfect one, because a limited amount of changes at the amino acid level is permissible as different amino acid sequences can result in very compatible physico-chemical properties. In spite of its indirectness and imperfect conservation, it gives rise to the simplest computational solution—multiple alignment [29]—to the problem of finding the active sites of an enzyme.

A multiple alignment can be thought of as a way of writing two or more sequences across the page. Some gaps may be inserted into the sequences in such a way that the number of columns having characters that are identical or that are representing similar physico-chemical properties is maximized. The positions corresponding to these columns are called “conserved positions”. The most conserved positions in a multiple alignment are good candidates of active sites of the enzyme, provided the sequences used in the multiple alignment are from suitably diverged species. That is, the sequences should be sufficiently diverged so that enough mutations have accumulated in positions that do not correspond to active sites. At the same time, the sequences should not be so wildly diverged that they no longer have the required enzyme function. Figure 2 shows a multiple alignment of several protein tyrosine phosphatase sequences. The candidate active sites are the conserved consecutive positions indicated by “*” and “.”.

3 A Couple of Interesting Twists

An interesting twist in the tale of active sites is the problem of finding key mutations that cause a protein to reduce the efficiency of its function. Here, one of the ancestor proteins with a function f has a mutation in one of its active sites for function f . This mutation reduces the efficiency of the protein. The

```

gi|126467|      FHFTSWPFDGVPFPTIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDANLD
gi|2499753|     FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGTCYIVIDINLD
gi|462550|      YHYTQWPDHGVPYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSNLQ
gi|2499751|     FHFTSWPDHGVPDITDILLINFRYLVRDYMKQSPESPILVHCSAGVGRTGTFIAIDRLIY
gi|1709906|     FQFTAWPDHGVPENPTPFLAFLRRVKTCLNP--PDAGPMVVHCSAGVGRTGCFIVIDANLE
gi|1264711|     LHFTSWPFDGVPFPTIGMLKFLKKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDANHA
gi|548626|      FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGTCYIVIDINLD
gi|131570|      FHFTGWPDHGVPYHATGLLGFVRQVKSASP--FNAGPLVVHCSAGAGRTGCFIVIDINLD
gi|2144715|     FHFTSWPDHGVPDITDILLINFRYLVRDYMKQSPESPILVHCSAGVGRTGTFIAIDRLIY
               ..* *** **      . *               ..***** ***. ** ..

```

Fig. 2. A snapshot of a multiple alignment of several protein sequences.

mutation is passed to a group of descendant proteins with function f at a lower efficiency, and becomes an invariant of this group.

Thus, to find key mutations that reduce the efficiency of a protein for function f , we proceed as illustrated in Figure 3. We first identify a group D_1 of proteins having function f at the normal level of efficiency. Then we identify a group D_2 of proteins having function f at the reduced level of efficiency. Then we identify a common active site in two groups of proteins so that two different invariants—one for each of the groups—are observed at the site. That is, the change in efficiency is traced to mutations in specific active sites in the first group which are inherited and conserved in the second group. This takes us from the concept of invariants to the concept of emerging patterns—patterns which are invariant in one group and are changed in a contrast group [12, 8]. A beautiful illustration of this logical solution can be found in the study of protein tyrosine phosphatases [15].

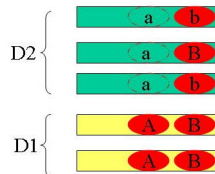


Fig. 3. The site “B” is consistently conserved in the D_1 group of sequences, but is not consistently mutated in the D_2 group. It is thus not a likely cause of D_2 ’s reduced efficiency; otherwise, the second sequence in the D_2 group which has an unmutated site “B” should have normal efficiency. The site “A” is consistently conserved in the D_1 group, and is consistently mutated in the D_2 group. Thus it is a possible cause of D_2 ’s reduced efficiency.

An important invariant of mutations underlies the twist in the tale of active sites above: Mutations are cumulative. That is, a mutation is passed on to future generations unless there is another mutation at the same site that replaces it. This invariant can be exploited in problems concerning the origin of species. The

human mitochondrial control region accumulates about 1 mutation every 10,000 years [27]. Given the short length of human history, the length of the mitochondrial control region, and each position in it has an equal chance to mutate, it is reasonable to assume that any position has a negligible likelihood of being mutated twice. In other words, a mutation in the mitochondrial control region that is observed in all instances of the ancestor species must also be observed in descendant species. Thus a link from an ancestor species to its descendant species can be traced.

A beautiful illustration of this idea can be found in the story of the origin of Polynesians [27], depicted in Figure 4. All indigenous Taiwanese have two mutations referred to as #189 and #217 in their mitochondrial control region. Indigenous Solomon Islanders have mutations #189, #217, and #261. Thus, we conclude that an indigenous Taiwanese or his descendant with the #261 mutation somehow travelled to the Solomon Islands, and all indigenous Solomon Islanders are his descendants. All Rarotongans have mutations #189, #217, #261, and #247. Similarly, we infer that a Solomon Islander or his descendant with the #247 mutation somehow reached Rarotonga, and present-day Rarotongans are his descendants.

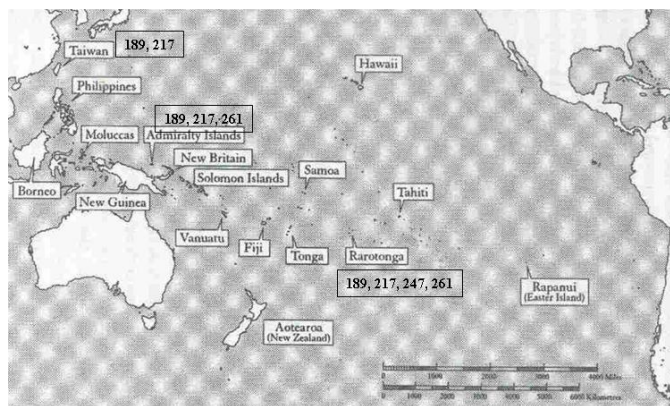


Fig. 4. Origin of Polynesian. Image credit: Sykes [27].

4 Invariants in Protein Function

There are two main invariants that determine the function of a protein: The three-dimensional conformation of the protein and the environment the protein is in. These invariants impose important constraints on the amino acid sequence of protein. For example, mutations in the sequence may completely change the

three-dimensional conformation of the sequence. Thus the sequence of the protein is also under pressure to be invariant. However, this invariant is indirect and does not have to be perfect. For example, a limited amount of changes at the amino acid level is permissible without severely affecting the three-dimensional conformation of the protein. Nevertheless, one can perform an abductive inference to predict that two proteins that exhibit a high level of sequence similarity are likely to have the same or similar function. This is the so-called “guilt by association” of similarity of sequences, exemplified by the classic paper of Doolittle and others [9].

The procedure of “guilt by association” is depicted in Figure 5. We compare the sequence of the unknown protein T with a database of protein sequences with known functions. Those proteins in the database that have high sequence identities or sequence alignment scores when compared to T are predicted to be homologs of T ; and T is predicted to have functions identical or similar to those of these homologous proteins. A pairwise alignment algorithm [20, 25] should be used for sensitive search of homologs. Due to the rapid increase in sequence database sizes, it is also common to sacrifice some amount of sensitivity in favour of significantly increased speed by first using short perfect matches to select likely candidate sequences before performing pairwise alignments [1].

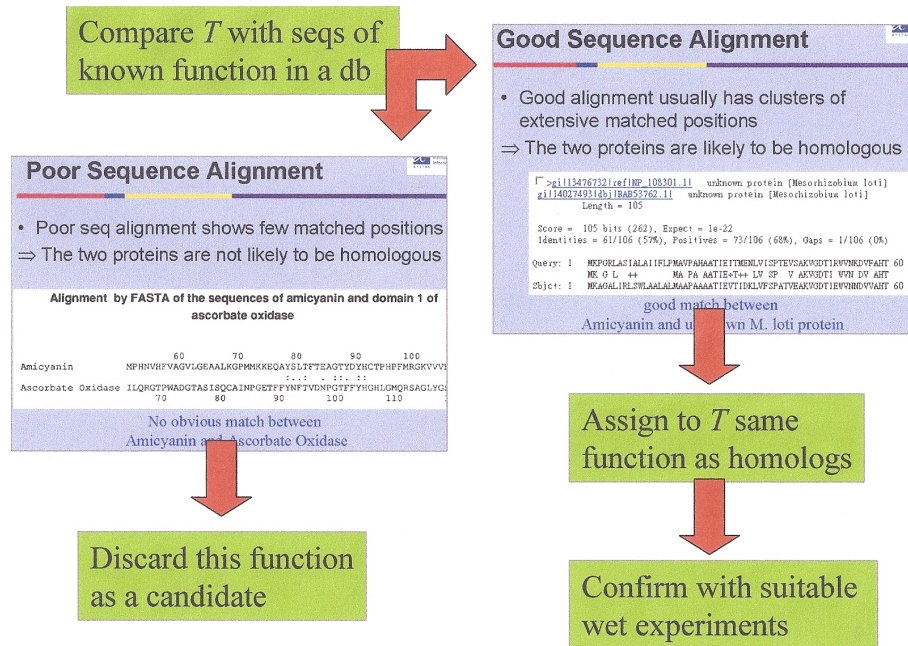


Fig. 5. Protein function prediction using “guilt by association” of sequence similarity.

However, there are many protein sequences that have very low sequence similarity to all proteins of known functions. In such a situation, we have to appeal to additional consequences of the two invariants of three-dimensional conformation and operating environment required for a protein function. I describe one such consequence below.

The invariant on a protein sequence, though indirect and imperfect, has an interesting and subtle consequence. Proteins exhibiting a function f , proteins exhibiting a different function g , and proteins exhibiting a function h have different three-dimensional conformations and possibly operate in different environments. So the sequences of these three groups of proteins have distinct invariant compositional characteristics. However, the differences of the invariant compositional characteristics of any two proteins of functions f and g are very likely to be very similar to the differences of the invariant compositional characteristics of any other two proteins of functions f and g ! On the other hand, these differences are very likely to be very different from the differences of the invariant compositional characteristics of two proteins of functions f and h , or of functions g and h .

In short, the differences of the invariant characteristics of one group of proteins compared to another group are also invariant, and are emerging patterns when contrasted with the differences compared to a third group. This logic is best illustrated by the comparison of apples to oranges and bananas in Figure 6, where the fruit X is deduced as an apple because its differences with orange₁, banana₁, and other fruits are identical to that of apple₁.

	orange ₁	banana ₁	...
apple ₁	color=red vs orange skin =smooth vs rough shape=round vs round	color=red vs yellow skin=smooth vs smooth shape=round vs oblong
orange ₂	color=orange vs orange skin =rough vs rough shape=round vs round	color=orange vs yellow skin=rough vs smooth shape=round vs oblong
fruit X	color=red vs orange skin =smooth vs rough shape=round vs round	color=red vs yellow skin=smooth vs smooth shape=round vs oblong
...

Fig. 6. Comparing apples vs oranges vs bananas. The fruit X is likely to be an apple because its differences with orange₁, banana₁, etc. are identical to that of apple₁.

To wit, we can associate two proteins as having the same or similar function by the similarity of the differences of their sequences compared to all other sequences. This is precisely the strategy followed by SVM Pairwise [14]. Here, a feature vector is generated for each protein by recording its pairwise alignment

score with each sequence in the database. To create a classifier for distinguishing proteins of function f from the rest, the feature vectors are divided into f vs non- f , and a support vector machine classifier is then trained. Given a new unknown protein, a feature vector is first generated by recording its pairwise alignment score with each sequence in the database. The feature vector is then given to the classifier for prediction. SVM Pairwise has much greater sensitivity and precision than the more direct guilt by association of sequence similarity described earlier. SVM Pairwise succeeds for two main reasons. Guilt by association of sequence similarity cannot be applied if a sequence has low similarity with the database and it does not make use of contrast groups. In contrast, SVM Pairwise does not care about the level of sequence similarity, so long as the sequence alignment scores have consistent differences between f vs non- f .

5 Invariants in Diseases

One of the popular problems in bioinformatics is the analysis of gene expression profiles for disease subtype diagnosis. Each disease and its various subtypes have their underlying causes. The causes are often difficult to decipher due to the complexity of molecular circuitries and gene-environment interactions. Nevertheless, different causes have different invariant down-stream effects that are useful as diagnostic indicators. These invariant down-stream effects are often—but not always—manifested as consistent gene expression profile differences in a large number of target genes over the different disease subtypes.

This type of invariant down-stream effects can be discovered in a variety of ways [18]. For example, in an unsupervised setting, one discards those genes with low variants, performs a bi-clustering of the remaining genes vs patient samples, and identifies the invariant gene expression profiles for each disease subtype. As another example, in a supervised setting, one groups the patient samples based on disease subtypes, computes a test statistics such as χ^2 for each gene to determine how well it separates one disease subtype from the rest, and identifies those genes that best distinguishes a subtype. Figure 7 is a beautiful illustration based on the gene expression profiles of childhood acute lymphoblastic leukemia samples [33].

Childhood acute lymphoblastic leukaemia (ALL) is the most common form of childhood cancer. It has as many as 6 different subtypes with differing treatment outcome. To avoid under-treatment, which causes relapse and eventual death, or over-treatment, which causes severe long-term side effects, accurate diagnostic subgroup must be assigned upfront so that the correct intensity of therapy can be delivered to ensure that the child is accorded the highest chance for cure [22]. Contemporary approaches to the diagnosis of childhood ALL require an extensive range of procedures including morphology, immunophenotyping, cytogenetics, and molecular diagnostics [22]. Such a multi-specialist expertise requirement is generally unsatisfiable in developing countries. Thus, even though childhood ALL is a great success story of modern cancer therapy with survival rates of

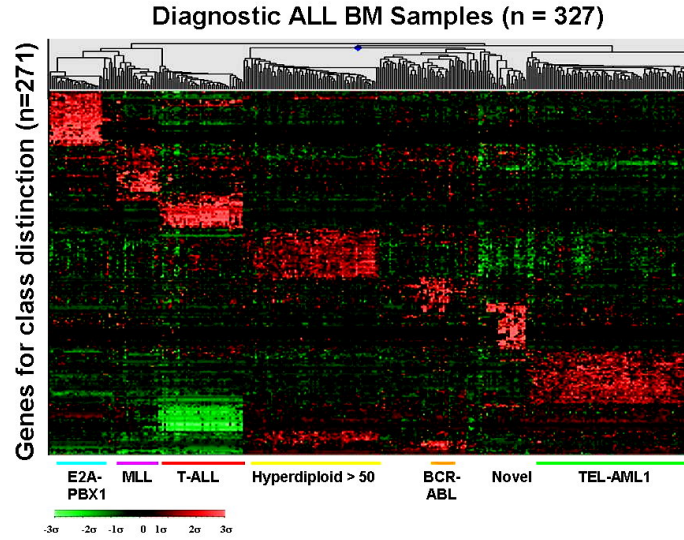


Fig. 7. Gene expression profiles of childhood ALL. Each row is gene. Each column is a patient. Image credit: Yeoh and others [27].

75–80% in major advanced hospitals [23], it is still a fatal disease in developing countries with survival rates of 5–20%.

Our microarray gene expression profiling followed by computational analyses described above accurately identifies each of the known clinically important subgroups of childhood ALL [33]. We achieve an exceedingly accurate overall diagnostic accuracy of 96% in a blinded test set illustrating the robustness of the invariants identified.

It is worth noting that about 2000 new cases of childhood ALL are diagnosed in ASEAN countries each year. About 50% of these cases need low-intensity therapy, 40% need intermediate intensity, and 10% need high intensity. This is a disease with a cure rate of >75% in Singapore. But in ASEAN countries, except Malaysia and Singapore, childhood ALL patients have a dismal 5–20% cure rates. This is mainly due to these countries' inability to deliver the correct intensity of therapy. Treatment for childhood ALL over 2 years for intermediate-risk costs US\$60k, good-risk costs US\$36k, and high-risk costs US\$72k. Treatment for relapse cases costs US\$150k. As the less developed ASEAN countries generally lack the ability to diagnose the subtypes of their childhood ALL patients, the treatment for intermediate risk case is conventionally applied for everyone, as it maximizes the expected benefit in such a situation; see Figure 8. If our single-test platform becomes broadly available, they can then adopt a more accurate risk-stratified treatment strategy. As shown in Figure 8, this can result in savings

of US\$52M a year yet with better cure rates and much reduced side effects, as the correct intensity of therapy is applied upfront.

Treatment	Cost–new cases	Cost–relapses	Total cost
Low-intensity treatment for everyone	\$36K * 2000	\$150K * 1000	\$222M
Intermediate-intensity treatment for everyone	\$60K * 2000	\$150K * 200	\$150M and 50% of patients have side effects
High-intensity treatment for everyone	\$72K * 2000	\$0	\$144M and 90% of patients have side effects
Risk-stratified treatment; viz., low intensity to 50%, intermediate intensity to 40%, high intensity to 10%	\$36K * 1000 + \$60K * 800 + \$72K * 200	\$0	\$98M

Fig. 8. Costs of treatment options for childhood ALL in ASEAN countries.

6 Remarks

Let me now summarize the key learnings of this tutorial.

I have considered several common bioinformatics applications such as recognizing active sites and key mutations, determining the origin of Polynesians, predicting the function of proteins, and diagnosing a disease and optimizing its treatment. We have seen that there are invariants underlying these problems, and the exploitation of such invariants and/or their consequences yield logical solutions to these problems.

I have used three paradigms in the exploitation of invariants here. The first paradigm is a direct search of an invariant in a group. An example of such a direct search is the application of finding active sites with the use of a multiple alignment algorithm. The second paradigm is a search of “emerging patterns”, where we look for patterns that are invariant in one group but are changed in a contrast or control group. The use of a contrast group helps isolate invariants that are fundamental to the target group, as opposed to invariants that are observed in a general population. An example of a search for emerging patterns is the application of finding key mutations that cause a group of proteins to reduce the efficiency in their function. The third paradigm is the concept of “guilt by

association”, where we deduce that two objects belong to the same type if they exhibit specific common invariants associated with that type. An example of this is the inference of protein function. These three paradigms are also used in combination. An example is the identification of gene expression profiles for diagnosing childhood ALL subtypes. Here, we look for gene expression profiles as emerging patterns that distinguish one ALL subtype from the other subtypes, and use such gene expression profiles to classify patients into the associated ALL subtypes.

I have also discussed the softer but still very important aspect of economics of bioinformatics. This is illustrated in the treatment optimization of childhood ALL. In particular, we have explained why the intermediate-intensity treatment is conventionally applied if the ALL subtype cannot be applied, and why a risk-stratified treatment based on bioinformatics analysis is a superior strategy.

I have briefly mentioned four kinds of computational techniques here. The first kind is that of multiple sequence alignment, where we determine how to best match up several sequences, as illustrated in the application of finding active sites. The second kind is that of sequence comparison, where we determine if two sequences are sufficiently similar, as illustrated in the application of protein function inference. The third and fourth kinds are those of statistical testing and machine learning, as illustrated in the analysis of childhood ALL gene expression data.

Paper length constraints do not allow a more detailed exposition of the above. The reader is encouraged to consult the following articles and references therein for more information. In particular, for sequence comparison, Waterman [30] provides an excellent theoretical background, Gusfield [10] provides an excellent algorithmic background, and Li et. al. [13] present the exciting recent development of using spaced seeds for extremely sensitive and efficient sequence comparison. For multiple sequence alignment, Thompson et. al. [29, 28] describe one of the most popular multiple alignment tool packages, and Chin et. al. [6] present a recent improvement in efficient multiple sequence alignment with performance guarantee. For protein function prediction, Altschul et. al. [1] describe the extremely popular BLAST approach to guilt by association of sequence similarity, Bateman et. al. [2] describe guilt by association of domain similarity as embodied in PFAM domains, Liao and Noble [14] describe guilt by association of similarity of dissimilarities as embodied in SVM Pairwise, Wu et. al. [32] describe guilt by association of similarity of phylogenetic profiles, Ma et. al. [17] describe guilt by association of secondary structures, Kung et. al. [11] describe guilt by association of similarity in gene expression profiles, and Chua et. al. [7] present the exciting recent development of guilt by association of similarity of interaction partners. For gene expression analysis, Slonim et. al. [24] is the classic paper that started the field, Miller et. al. [18] is an excellent overview of the issues and techniques, Broberg [4] is a good discussion on several popular test statistics, Breitling and Herzyk [3] describe new rank-based test statistics, Nijima and Kuhara [21] describe new kernel subspace methods for multiclass classification,

and Subramanian et. al. [26] present the exciting recent development of the gene set enrichment analysis approach.

Acknowledgements

I would like to thank Prof. Bruno Buchberger and the organizers of AB2007 for inviting me to present this tutorial paper.

References

1. S. F. Altschul, T. L. Madden, et. al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
2. A. Bateman, E. Birney, et. al. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Research*, 27(1):260–262, 1999.
3. R. Breitling and P. Herzyk. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *Journal of Bioinformatics and Computational Biology*, 3(5):1171–1190, 2005.
4. P. Broberg. Statistical methods for ranking differentially expressed genes. *Genome Biology*, 4:R41.1–R41.9, 2003.
5. D. G. Brown, M. Li, and B. Ma. A tutorial of recent developments in the seedings of local alignment. *Journal of Bioinformatics and Computational Biology*, 2(4):819–842, 2004.
6. F. Y. L. Chin, N. L. Ho, et. al. Efficient constrained multiple sequence alignment with performance guarantee. *Journal of Bioinformatics and Computational Biology*, 3(1):1–18, 2005.
7. H. N. Chua, W.-K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22:1623–1630, 2006.
8. G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proc. 5th ACM SIGKDD Intl Conf on Knowledge Discovery & Data Mining*, pages 15–18, San Diego, 1999.
9. R. F. Doolittle, M. W. Hunkapiller, et. al. Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science*, 221:275–277, 1983.
10. D. Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, 1997.
11. S.-Y. Kung, M.-W. Mak, and I. Tagkopoulos. Symmetric and asymmetric multi-modality biclustering analysis for microarray data matrix. *Journal of Bioinformatics and Computational Biology*, 4(2):275–298, 2006.
12. J. Li and L. Wong. Identifying good diagnostic genes or genes groups from gene expression data by using the concept of emerging patterns. *Bioinformatics*, 18:725–734, 2002.
13. M. Li, B. Ma, et. al. PatternHunter II: Highly sensitive and fast homology search. *Journal of Bioinformatics and Computational Biology*, 2(3):417–440, 2004.
14. L. Liao and W. S. Noble. Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In *Proc. 6th Annual Intl Conf on Research in Computational Molecular Biology*, pages 225–232, 2002.

15. K. L. Lim, P. R. Kolatkar, et. al. Interconversion of kinetic identities of the tandem catalytic domains of receptor-like protein-tyrosine phosphatase PTP- α by two point mutations is synergistic and substrate-dependent. *Journal of Biological Chemistry*, 273(44):28986–28993, 1998.
16. H. Liu and L. Wong. Data mining tools for biological sequences. *Journal of Bioinformatics and Computational Biology*, 1(1):139–168, 2003.
17. B. Ma, L. Wu, and K. Zhang. Improving the sensitivity and specificity of protein homology search by incorporating predicted secondary structures. *Journal of Bioinformatic and Computational Biology*, 4(3):709–720, 2006.
18. L. D. Miller, P. M. Long, et. al. Optimal gene expression analysis by microarrays. *Cancer Cell*, 2:353–361, 2002.
19. S. Mukherjee and S. Mitra. Hidden Markov models, grammars, and biology: A tutorial. *Journal of Bioinformatics and Computational Biology*, 3(2):491–526, 2005.
20. S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:444–453, 1970.
21. S. Nijijima and S. Kuhara. Multiclass molecular cancer classification by kernel subspace methods with effective kernel parameter selection. *Journal of Bioinformatics and Computational Biology*, 3(5):1071–1088, 2005.
22. C. H. Pui and W. E. Evans. Acute lymphoblastic leukemia. *New England Journal of Medicine*, 339:605–615, 1998.
23. M. Schrappe, A. Reiter, et. al. Improved outcome in childhood acute lymphoblastic leukemia despite reduced use of anthracyclines and cranial radiotherapy: Results of trial ALL-BFM 90. *Blood*, 95:3310–3322, 2000.
24. D. K. Slonim, P. Tamayo, et. al. Class prediction and discovery using gene expression data. In *Proc. 4th Intl Conf on Computational Molecular Biology*, pages 262–271, 2000.
25. T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.
26. A. Subramanian, P. Tamayo, et. al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci. USA*, 102(43):15545–15550, 2005.
27. B. Sykes. *The Seven Daughters of Eve*. Gorgi Books, 2002.
28. J. D. Thompson, T. J. Gibson, et. al. The CLUSTAL-X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, 25(24):4876–4882, 1997.
29. J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.
30. M. S. Waterman. *Introduction to Computational Biology: Maps, Sequences, and Genomes*. CRC Press, 2000.
31. J. C. Wooley and H. S. Lin, editors. *Catalyzing Inquiry at the Interface of Computing and Biology*. National Academy Press, 2005.
32. J. Wu, S. Kasif, and C. DeLisi. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, 19(12):1524–1530, 2003.
33. E.-J. Yeoh, M. E. Ross, et. al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.