AN APPLICATION OF ASSOCIATION RULE MINING TO HLA-A*0201 EPITOPE PREDICTION

TOM MILLEDGE, GAOLIN ZHENG, GIRI NARASIMHAN*
School of Computer Science
Florida International University
Miami, FL 33199

This paper presents a novel approach to epitope prediction based on the clustering of known T-cell epitopes for a given MHC class I allele (HLA-A*0201). A combination of association rules (ARs) and sequence-structure patterns (SSPs) was used to do the clustering of training set epitopes from the Antijen database. A regression model was then built from each cluster and a peptide from the test set was declared to be an epitope only if one or more of the models gave a positive prediction. The sensitivity (TP/TP+FN) of the AR/SSP regression models approach was higher than that of a single regression model built on the entire training set, and was also higher than the sensitivity measures for SYFPEITHI, Rankpep, and ProPred1 on the same test set.

Keywords: Data Mining, regression, epitopes, prediction, Sequence-Structure patterns.

1. Introduction

The major histocompatibility complex (MHC) molecules are cell-surface glycoproteins that bind peptide fragments derived from intracellular and extracellular proteins. MHC proteins are divided into two major subfamilies, class I and class II. Each allelic variant of a class I MHC molecule selects binding peptides based upon the complementary structure of the peptide and the polymorphic pockets within the MHC binding groove. These peptides, usually about 8 to 10 residues in length, are derived from cytosolic proteins (which are cleaved by proteosomes), and are translocated to the endoplasmic reticulum by peptide transporters associated with antigen processing before final association with MHC class I molecules. The peptide-MHC complexes are then transported to the cell membrane where they can be recognized by CD8+ T cells. MHC-bound peptides that are recognized by T-cells are referred to as *epitopes*.

Structurally, the N- and C- termini of the bound peptide are buried in the binding groove of the MHC class I molecule and the central peptide region bulges slightly outward. Polymorphic side-chains define six pockets (A-F) in the MHC molecule which interact with some of the peptide side-chains [1, 2]. Typically, most of the epitopes associated with a particular MHC class I allele (such as the commonly occurring and widely studied HLA-A*0201 allele) are

* Corresponding author

related by sequence and share characteristic structural motifs [3, 4]. In HLA-A*0201 MHC /peptide complexes, the binding of hydrophobic residues between peptide position 2 (P2) and HLA-A*0201 pocket B and between the peptide C-terminus (P9 for nonamer peptides) and pocket F are the largest contributors to the high-affinity binding. These two positions (the "anchors") have been found to be necessary, but not sufficient, as the predictions based solely on the residues at these positions are only about 30% accurate [5]. Due to the imprecision of sequence binding motifs [6], a number of profile-based methods have been developed for epitope prediction, such as SYFPEITHI [7], Rankpep [8], and ProPred1 [9]. The predictive ability of all these single profile methods is limited by their critical dependence on the assumption that each amino acid acts independently to generate positive or negative effects to binding energy [10]. Despite their short length and relatively constrained termini, substitutions in the center of a peptide bound to class I MHC proteins have frequently been found to affect the positions of all of the residues within the peptide [11, 12]. Furthermore, the concept of position is problematic when comparing the binding patterns of different length peptides. As a result, profile methods for epitope prediction typically use separate position-specific scoring matrices (PSSMs) for epitopes of different lengths. In this paper, we extend the use of separate models by clustering MHC class I epitopes of a given length (nine) based on evidence of correlated substitutions within each putative cluster.

What substitutions in the epitopes are correlated? In order to discover and characterize the nature of the interactions between substituted residues in active (i.e., immunogenic or functional) MHC/peptide complexes, a study was undertaken here of the patterns of association rules (ARs) found by applying the Apriori data mining algorithm to all known T-cell epitopes in the Antijen (Jenpep) database for epitopes of length nine (nonamers) that are bound by the HLA-A*0201 allele. In the more general problem of protein function prediction, the use of association rule mining has shown promise in correlating protein sequence, structure and function [13]. The Apriori algorithm used here is an approach for the discovery of association rules in large sets of items ('itemsets') and was originally developed for transactional data [14, 15]. The goal of Apriori is to find all itemsets, i.e., frequent subsets, that have transaction support above a specified threshold. In this context, the itemset is a specific set of residues at fixed positions within the peptide sequence, and the support is the percentage of peptides in the database that contain the given residues at the specific positions. To obtain associations between positions in the peptide, the Apriori association rule mining attempts to find associations for all residue

itemsets. An association rule is an implication of the form $X_i \rightarrow Y_j$[s, c], where X and Y are residues, i and j are positions in the peptide, s% is the *support* of the rule and c% is the *confidence*. For example, the rule Gly5→Gly3[10.1, 36.4] states that of the 10.1% of the time that Glycine occurs at the fifth position of the peptide, a Glycine is also found at the third position 36.4% of the time. Through the discovery of association rules in the training set of nonamer peptides that are known to bind the HLA-A*0201, we have attempted to find sets of residues that frequently occur together, so as to enable the generation of sequence patterns from the ARs. Epitopes matching these sequence patterns were then grouped into clusters according to their conformance with observed patterns of sequence and structure derived from a structural analysis of all known HLA-A*0201/peptide complexes in the Protein Data Bank (PDB). We hypothesize that it is only within these structurally related clusters of known T-cell epitopes that independent substitutions of residues occur. To test this premise, we generated one regression model from each of these clusters and used them for epitope prediction against a separate test set. We then compared the results from this ensemble method with the performance of a single regression model generated from the unpartitioned training set.

2. Procedures

2.1 Structure Analysis

In a previous work, we showed how to extract patterns with both sequence and structure components (sequence-structure patterns, or SSPs) from a set of related PDB structures [16]. In order to characterize and classify the observed conformations of nonamer peptides bound to the MHC class I HLA-A*0201 allele, we applied this method to twenty-five HLA-A*0201 MHC/peptide complexes from the PDB. The RMSD was measured between each pair of protein complexes using only the Cα atoms of those MHC residues which came into contact with the peptide backbone. The complex with the smallest average RMSD (1eey) to all the other complexes was then used as a structural template for the alignment of the other MHC complexes. Pocket B is the binding site for relatively small, branched aliphatic residues at the peptide anchor position P2. Pocket D is formed by residues Tyr99, His114, Leu156 and Tyr159. Although other HLA alleles select for basic or acidic residues due to the presence of either an Asp156 or Arg156, respectively, the uncharged Leu156 of HLA-A*0201 allows a large variety of side-chain binding configurations. Examples of the variety of peptide backbone and sidechain conformations allowed by the HLA-A*0201 allele are illustrated in Figure 1.

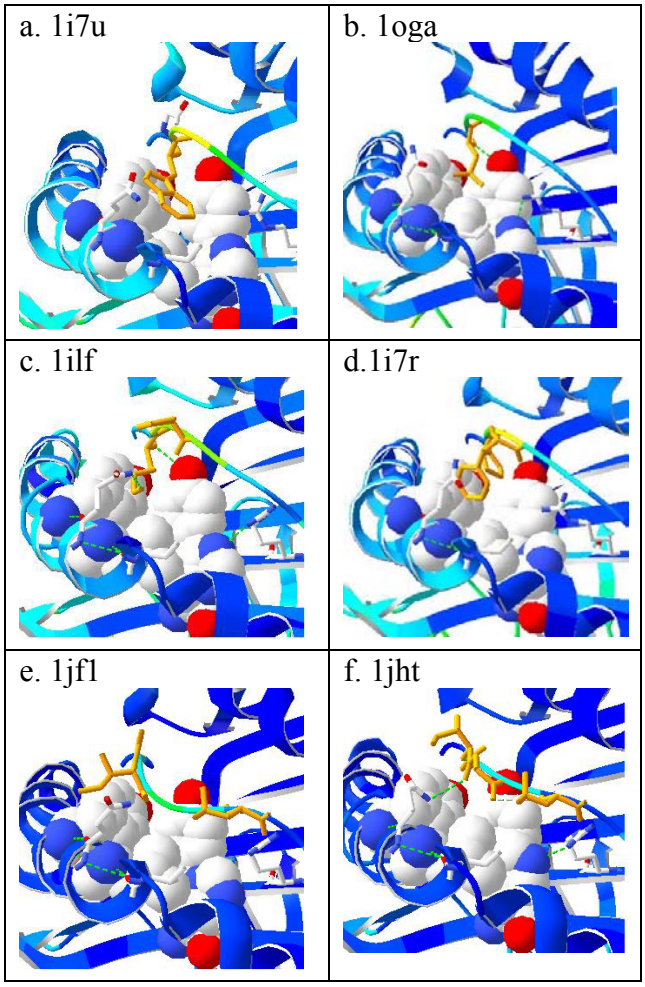

Figure 1a-f: A variety of observed MHC/peptide conformations at the interface between peptide positions P3-P5 and HLA-A*0201 pocket D. Pocket D residues sidechains are shown as space-filled molecules. The MHC binding groove is shown between the two α-helices and running from the background (upper left) to the foreground (lower right) in all the figures.

Among the MHC/peptide complexes in the PDB, the majority of peptide residues at position P3 have large neutral side-chains, such as Phe, Trp and Leu, which lie flat across the pocket, parallel to the base of the MHC binding groove (Figures 1a. and 1b.). Another allowable conformation is to have a Lys at P3 span pocket D and make a salt bridge with Gln155 (Figure 1c.). In some cases

Pocket D is filled by a peptide residue from positions P4 or P5 (Figures 1d. and 1e.). When P3 is a Gly residue, Gln155 is able to make a hydrogen bond with the backbone (Figure 1f.). The conformation of 9-mers 1i7u, 1ily, 1i7r and 1b0g allows the formation of a hydrogen bond between the backbone at peptide position P3 and His70.

Although the backbone of peptides bound to MHC class I alleles bow outward in the middle, the side-chain of the residue at either positions P5 or P6 is usually oriented downward to interact with the peptide-binding groove. In HLA-A*0201, the large residues Arg97 and Phe9 fill pocket C, such that the side-chains from P5/6 must bend toward pocket D. Gln155 is also able to form bonds with charged residues at P5/6, thereby allowing a wide variation in the allowable residues in this range of the peptide. The structures 1oga and 1hhi have a hydrogen bond between the Thr sidechain at P8 and Lys146. At the C-terminal position of the bound peptide, a network of conserved residues at pocket F stabilizes the MHC/peptide complex. The Tyr116 of HLA-A*0201 restricts entrance to the hydrophobic pocket and selects strongly for the singly branched aliphatic residues Leu or Val.

By comparing the alignments of the peptide backbones and the patterns of hydrogen bond formation, eight binding peptides SSPs were identified (Table 1). Within these structural clusters, the pairwise RMSD of the Cα atoms was less than 0.50 angstroms. These SSPs were used later to guide the clustering of association rules (ARs) which were generated as described below.

Table 1: HLA-A*0201-bound peptide SSPs

Sequence-structure pattern (SSP)	Structures	Addl. H-bonds
[FLY]L[SK][AE][PL]V[GH][GI][VL]	1eey, 1ily, 1eez	Position 6/ Thr73
GILGFVFTL	1oga, 1hhi	Position 8/ Lys146
[FIY]LKEPVHGV	1ily, 1akj, 1hhj, 1ilf	Position 7/ Gln155
[FA][AL][PW]G[FV][FV]P[VY]L	1i7u, 1i7r, 1b0g, 1i7t	Position 3/ His70
LLFGY[AP][VR][AY]V	1qrn, 1qsf, 1qse, 1bd2, 1a07	None
[LM]L[FL][GS][VY]P[LV][LY][VL]	1duz, 1hhk, 2clr, 1im3	None
TLTSCNTSV	1hhg	Position 8/ Asp77
ALGIGILTV	1jht	None
IISAVVGIL	1qr1	None

2.2 Association Rule (AR) Mining

In our initial attempts to use association rule mining on the MHCPEP database [17] and on a database of known HLA-A*0201-binding peptides published by Doytchnova [18], we discovered that virtually all of the highest confidence rules resulted from relatively small sets of proteins in which only one or two residues were changed by the original researcher and the remainder were held constant as a template. In general, these rules had little or no correspondence to known patterns of peptide structure or residue interaction. Subsequent studies with the larger Antijen (formerly JenPep) database [19] resulted in a number of rules which did correspond to sequence motifs generated by a structure analysis of known HLA-A*0201/peptide complexes that had been deposited in the Protein Data Bank (PDB). From these exploratory studies, a complete analysis of patterns of association between the residues of known epitopes of length nine of the HLA-A*0201 allele was undertaken.

To avoid the generation of biased ARs of high confidence, a number of data preprocessing steps were required. Although, in theory, there was no problem with duplicate transactions, in practice the determination that identical peptides were truly discovered independently could not always be completely established. Therefore, it was decided that duplicate peptides were only to be counted once, even in the case where they had been discovered independently from separate genomes. A more serious problem was detecting the presence of peptides generated by residue substitution studies. The only way to resolve this issue was to track the reference or references provided by the Antijen database for each residue, and read the methodologies of the original experiments. Since this step was required for each of the 400 candidate members of the data set, the following procedure was established:

1. Epitopes discovered from scans of either actual or putative genes from genomic DNA were retained in the training set.
2. If the originating research was primarily a study of MHC-binding, in which the candidates were synthesized according to an experimental procedure, then these peptides which were determined to be immunogenically active (epitopes) were moved from the training set to the test set.

This procedure left 250 naturally occurring nonamer peptides in the training set and 71 artificial known epitope nonamer peptides in the test set. Ideally, training set epitopes would only be those whose discovery resulted from

completely unsupervised scans of all possible peptides from the source genome. However in practice, almost all workers in this area utilized one or more allele-specific sequence motifs to narrow their search. In the case of the HLA-A*0201 nonamers under study here, these motifs were often some variation of x-[LIM]-x(6)-[LV], where the second residue was either Leu, Ile or Met, and the last residue was either Leu or Val. As a result of this unavoidable bias built into the database, discovered associations between the second and ninth positions had to be rejected as artifacts of the candidate generation process, as discussed below.

The Apriori implementation by Christian Borgelt [20] was used for the association rule discovery. Initially, both the confidence and the support levels were set to 10% in order to generate as many rules as possible. The low support was also necessary to detect faint associations between small groups of residences (referred to variously as "substitution groups" or "equivalence classes") that are known to substitute for each other with high frequency [21]. For example if either I, L or V (at position 5, say) occurred with a W (say, at position 3) with a support of 40% and with high confidence, then the actual support of the rules I5→W3, L5→W3 and V5→W3 would occur with a support of some fraction of the 40% support for the group as a whole. Rules generated using the parameters above were then sorted by confidence level and analyzed. The minimum confidence level was subsequently raised to 20%, as the combination of (relatively) low support and confidence below 20% resulted in matching itemsets that were too small to be used for clustering purposes. Using a support parameter of 10% and a confidence parameter of 20%, a total of 80 candidate two-way association rules were generated. As discussed previously, the candidate epitopes contained a known bias between the residues in positions two and positions nine. As a result, all association rules between these positions were discarded, irrespective of confidence. The remaining question was how to interpret rules between specific residues at one of the anchor positions and specific residues at other positions of the peptide. It was decided that these rules could be retained if the support of this type of rule was significantly higher (50%) than the support of the anchor residue alone. Any other rule containing a residue known to occur alone at support level of over 10% (that is, irrespective of any associations with other residues) was also subjected to this increased support requirement. As a result of these procedures, 36 of the 80 candidate rules were discarded, leaving 44 association rules of interest.

2.2.1. Single Regression Model Approach

Regression models have been used previously for predicting binding affinity [10, 18]. In this paper, we built regression models based on assumption of independent binding where the response variable is either 1 for the epitopes or 0 for the non-epitopes (eq. 1).

$$Y = const + \sum_{i=1}^{9} p_i + \varepsilon \quad (1)$$

For control purposes, we built a single regression model that used all 250 peptides in the training set. A total of 250 non-epitopes (negative training cases) were generated randomly since a randomly generated peptide was very unlikely to be an epitope. We applied the model to the test data and obtained precision and sensitivity measures for the model at a threshold value of 0.5. A test peptide was declared to be an epitope by the program, if its predicted value was greater than the threshold.

2.2.2. AR/SSP Regression Models Approach

A total of 44 ARs of confidence greater than 20% were identified as described above. Each of these ARs were converted to a sequence pattern (AR sequence pattern) and used to search both the training set and the PDB sequences. For example, the association rule 7A→4G would give rise to the sequence pattern XXXGXXAXX, where the letters X represent an arbitrary residue. The results of these searches (the clusters) were merged based on the following criteria:

1. The AR sequence pattern matched one of the SSPs derived from the structural analysis.
2. The set of peptides matched by the AR sequence pattern overlapped with at least 25% of the peptides in an existing cluster.
3. One or more consensus residues existed at unconstrained positions within the cluster, which matched the corresponding positions of an existing cluster.

The first of these rules followed from the hypothesis that a specific peptide backbone structure would constrain the allowable residues at one or more positions, thereby creating a residue composition bias that should be detectable by association rule mining of a sufficiently large database of known epitopes. The second rule was formulated to avoid a large number of clusters that might be too small to allow the generation of a regression model. If there was an approximately 25% overlap between clusters that did not match PDB sequences

that were known to have separate structures, it was determined that these peptide clusters were likely to be structurally related, and could therefore be clustered together. The third rule above also arose from the need to consolidate the clusters. If a residue identity consensus emerged at the positions unconstrained by the matching sequence patterns, it was determined that these clusters were likely to be structurally related and were therefore merged.

A total of eleven clusters were generated by this method from 119 of the 250 training set epitopes which matched one or more of the 44 high confidence ARs. The largest of these clusters was 38 epitopes and the smallest was 5 epitopes. Five clusters from this group also corresponded to a previously identified SSP. The twelfth cluster consisted of the remaining 131 members of the training set which did not match either a SSP or an AR.

An example cluster is the set of sequences: AIIDPLIYA, AIIRILQQL, AIMDKNIIL, GIAGGLALL, GIGILTVIL, GILGFVFTL, GILTVSVAV, IISAVVGIL, KIFGSLAFL, LIGNESFAL, QILKGLLFL, RIIYDRKFL, RILQQLLFI, TILLGIFFL, VIYQYMDDL, and YIGEVLVSV which matched one or more of the association rules: 2I→9L[10.4%, 46.2%], 2I→3L[10.4%, 30.8%] and 2I→6L[10.4%,26.9%] The rule 2I→9L also matched the SSP IISAVVGIL (1qr1).

The peptides from each cluster were then paired with the non-epitope training cases, as described above, to use as training data. Twelve separate regression models were then built based on the training data sets. A testing peptide was declared to be an epitope only if one or more of the twelve regression models gave a positive prediction. For each model a value greater than the threshold of 0.5 was used to categorize the candidate peptide as an epitope, as was done for the single regression model above. Precision and sensitivity scores using the single regression model approach, the AR/SSP regression models approach, as well as for the SYFPEITHI, Rankpep and ProPred1 methods can be found below in Table 3.

3. Results

The results from the above experiments are summarized in Table 3. The performance of the ensemble of models constructed from AR/SSP clusters had a marginally higher sensitivity than the regression model constructed the training set as a whole. Although the precision values displayed by the other profile methods were all slightly higher than for the AR/SSP method, the more critical sensitivity values were significantly lower. Rankpep, in particularly showed poor results on the test set.

Table 3: Sensitivity and precision measures for the AR/SSP and Single Model methods compared with three other profile methods.

	Sensitivity	Precision
Single Model	93.1%	83.8%
AR/SSP Models	94.4%	81.9%
SYFPEITHI	77.8%	100%
Rankpep	13.9%	83.3%
ProPred1	70.8%	91.1%

In order to verify the independence of the test set, we also performed a 5-fold cross-validation using the single regression model using all known epitope samples (i.e., training set combined with test set). The 5-fold cross-validation error rate ((FP+FN)/(TP+TN+FP+FN)) was 5.4%, and the 5-fold cross-validation precision and sensitivity values were 94.3% and 95.0%, respectively. Since the figures for precision and sensitivity were even higher than with the test set alone, it supports the claim that the choices of training and test sets used in our experiments were relatively unbiased.

The rejection by SYFPEITHI of all the randomly generated peptides in the negative test set is strong evidence that none of these were true epitopes.

4. Discussion

In this paper, the term sequence-structure pattern (SSP) is given to refer to a set of proteins which contain both a common sequence motif and a structural template. If a model assumes that residues are allowed to substitute for one another, independent of substitutions at other positions in the protein, then a static backbone configuration is also a corollary assumption. By examining the peptide sequences associated with HLA-A*0201/peptide structures, we have attempted to identify novel sequence motifs that can be explicitly associated with a specific binding conformation. Since, by definition, independent substitution implies that there are no associations between the residue identities of distinct positions within the peptide, we have also utilized the technique of association rule mining to discover when this positional independence is violated. Although the initial results of this application of association rule mining to the epitope prediction problem are promising, a number of problems remain to be solved before this technique can be applied more broadly. In particular, this method is dependent on a large database of known epitopes for a

given allele. The HLA-A*0201 allele was chosen primarily because it is the most widely studied, and had the largest collection of both associated epitopes and of determined allele/peptide structures in the PDB. Clearly this prediction method would be more problematic using a smaller database which possibly could not be clustered effectively. A related problem is the supervised nature of the association rule clustering used here, which depended heavily on associations discovered independently through an analysis of known structures.

Although the very nature of the immune system and the degree of MHC polymorphism has proved to be an obstacle for the generation of sufficient MHC binding and T-cell epitope data for data mining approaches to epitope prediction to date [6], the growth of databases such as Antijen are sure to facilitate this type of data-driven approach in the future. By clustering peptides containing known sequence-structure patterns, association rules, or both, we have attempted to re-establish the prerequisite conditions under which independent substitution can be assumed.

Acknowledgements

Research of GN was supported in part by NIH Grant P01 DA15027-01.

References

1. Fremont, D., M. Matsumura, E. Stura, P. Peterson, and I. Wilson, *Science*, **257**, 5072 (1992).
2. Madden, D.R., D.N. Garboczi, and D.C. Wiley, *Cell*, **75**, 4 (1993).
3. Kast, W., R. Brandt, J. Sidney, J. Drijfhout, R. Kubo, H. Grey, C. Melief, and A. Sette, *J. Immunol.*, **152**, 8 (1994).
4. Zhang, C., A. Anderson, and C. DeLisi, *J Mol Biol*, **281**, 5 (1998).
5. Ruppert, J., J. Sidney, E. Celis, R. Kubo, H. Grey, and A. Sette, *Cell*, **74**, 5 (1993).
6. Flower, D.R., *Curr. Opin. Drug Discov. Devel.*, **6**, 3 (2003).
7. Rammensee, H.-G., J. Bachmann, N.P.N. Emmerich, O.A. Bachor, and S. Stevanovic, *Immunogenetics*, **50**, 3-4 (1999).
8. Reche, P.A., J.-P. Glutting, and E.L. Reinherz, *Hum. Immunol.*, **63**, 9 (2002).
9. Singh, H. and G.P.S. Raghava, *Bioinformatics*, **19**, 8 (2003).
10. Parker, K.C., M.A. Bednarek, and J.E. Coligan, *J. Immunol.*, **152**, 1 (1994).
11. Sharma, A.K., J.J. Kuhns, S. Yan, R.H. Friedline, B. Long, R. Tisch, and E.J. Collins, *J. Biol. Chem.*, **276**, 24 (2001).

12. Sliz, P., O. Michielin, J.-C. Cerottini, I. Luescher, P. Romero, M. Karplus, and D.C. Wiley, *J. Immunol.*, **167**, 6 (2001).
13. Satou, K., T. Ono, Y. Yamamura, E. Furuichi, S. Kuhara, and T. Takagi. in *Proc. Int. Conf. Intell. Syst. Mol. Biol.* (1997).
14. Agrawal, R., T. Imielinski, and A. Swami, *SIGMOD Record (ACM Special Interest Group on Management of Data)*, **22**, 2 (1993).
15. Agrawal, R., T. Imielinski, and A. Swami, *IEEE Transactions on Knowledge and Data Engineering*, **5**, 6 (1993).
16. Milledge, T., G. Zheng, and G. Narasimhan, *Proceedings of the Annual International Conference on Computational Molecular Biology, RECOMB*: Submitted 10/04. (2005).
17. Brusic, V., G. Rudy, A.P. Kyne, and L.C. Harrison, *Nucleic Acids Res.*, **26**, 1 (1998).
18. Doytchinova, I.A., M.J. Blythe, and D.R. Flower, *J. Proteome Res.*, **1**, 3 (2002).
19. McSparron, H., M.J. Blythe, C. Zygouri, I.A. Doytchinova, and D.R. Flower, *J. Chem. Inf. Comput. Sci.*, **43**, 4 (2003).
20. Borgelt, C. and R. Kruse. in *15th Conference on Computational Statistics*. (2002). Berlin, Germany.
21. Wu, T.D. and D.L. Brutlag. in *ISMB-96*. (1996).