BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 24 2011, pages 3415–3422
doi:10.1093/bioinformatics/btr594

Data and text mining

Advance Access publication October 28, 2011

Cascade detection for the extraction of localized sequence features; specificity results for HIV-1 protease and structure–function results for the Schellman loop

Nicholas E. Newell
21 Parkview Road, Reading, MA 01867, USA

Associate editor: Martin Bishop

ABSTRACT

Motivation: The extraction of the set of features most relevant to function from classified biological sequence sets is still a challenging problem. A central issue is the determination of expected counts for higher order features so that artifact features may be screened.

Results: Cascade detection (CD), a new algorithm for the extraction of localized features from sequence sets, is introduced. CD is a natural extension of the proportional modeling techniques used in contingency table analysis into the domain of feature detection. The algorithm is successfully tested on synthetic data and then applied to feature detection problems from two different domains to demonstrate its broad utility. An analysis of HIV-1 protease specificity reveals patterns of strong first-order features that group hydrophobic residues by side chain geometry and exhibit substantial symmetry about the cleavage site. Higher order results suggest that favorable cooperativity is weak by comparison and broadly distributed, but indicate possible synergies between negative charge and hydrophobicity in the substrate. Structure–function results for the Schellman loop, a helix-capping motif in proteins, contain strong first-order features and also show statistically significant cooperativities that provide new insights into the design of the motif. These include a new ‘hydrophobic staple’ and multiple amphipathic and electrostatic pair features. CD should prove useful not only for sequence analysis, but also for the detection of multifactor synergies in cross-classified data from clinical studies or other sources.

Availability: Windows XP/7 application and data files available at: https://sites.google.com/site/cascadedetect/home.

Contact: nacnewell@comcast.net

Supplementary Information: Supplementary information is available at *Bioinformatics* online.

Received on August 1, 2011; revised on October 20, 2011; accepted on October 23, 2011

1 INTRODUCTION

A localized sequence feature is defined here as a set of one or more sequence elements, such as amino acids, which occur at particular sequence positions in a set of equal-length sequences classified and aligned with respect to some biological function. Such a sequence set might represent a set of peptides recognized and cleaved by a particular protease, or a set of peptides that correspond to a particular structural motif in proteins. A ‘first-order’ feature specifies an element at a single position, whereas a ‘higher order’ feature specifies a set of two or more elements occurring together in a sequence at particular positions. A first-order feature that is significantly overrepresented in a sequence set may indicate the presence of an important physical effect in the relevant biological system involving the feature’s element, such as a preference for a negatively charged amino acid near a cleavage site. A significantly overrepresented higher order feature may represent an important cooperativity in the system, such as an electrostatic or hydrophobic interaction between two or more amino acid side chains.

Algorithms have been developed for feature selection in sequence sets which utilize correlation scans, artificial neural networks, genetic algorithms, support vector machines, decision trees, Markov methods and other techniques. For reviews, see Saeys *et al.* (2007) and Fogel (2008). A number of these techniques have been judged effective at the task of feature selection for sequence classification, but less progress has been made in the area of optimal feature detection for the purpose of analysis of the important effects in the data. A central problem which remains is the determination of the abundance that should be expected for a feature in a sequence set that may contain other significant features as well as random variation. If expected feature counts are not known, then it is not possible to determine whether an abundant feature represents an important effect, occurs by chance, or is an artifact which appears because it shares elements with the features that are responsible for the generation of function. Artifacts provide less than optimally accurate representations of the physical effects in the data: artifacts which are of lower order than the generator features which produce them can cause underestimation of complexity, whereas artifacts that combine elements of generator features masquerade as synergies. A method of screening out artifacts and accurately recovering generator features of all orders must be developed before structure–function datasets can contribute fully to accurate biochemical model building and efficient experimental design. Although researchers are now explicitly addressing the issue of dependencies in multivariate data (see e.g. Guyon *et al.*, 2002; Leek and Storey, 2008; Peng *et al.*, 2005; Yu and Liu, 2003), slow progress on the separation of higher order features from artifacts in sequence sets may have delayed experimental work directed toward the detection of cooperative interactions.

The present work introduces a new algorithm, cascade detection (CD), which computes expected feature counts, screens artifacts and identifies statistically significant localized features of all orders in sets of short sequences, or features of orders up to a specified maximum in longer sequences. The method detects first-order features that are either overrepresented or underrepresented and higher order features that are overrepresented. The algorithm

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

is tested on synthetic data, and then used to analyze amino acid sequence sets from problems in two different domains to demonstrate its broad utility: a study of HIV-1 protease (HIV PR) cleavage specificity applies the algorithm to a problem in which features play a transient sensing/recognition role, while an analysis of the Schellman loop motif in proteins evaluates features which are structural and therefore more permanent.

2 METHODS

2.1 The feature detection problem

As defined above, a localized feature is a set of sequence elements that occur together at particular, not necessarily contiguous sequence positions in a set of sequences. A variable-position feature may be identified as a group of features that differ only by position shifts, while a compound feature that allows multiple possible elements at individual positions may be identified as a set of related individual features. The feature detection problem is best illustrated by an analysis of synthetic data. Table 1 shows a list of 10 synthetic features of orders up to 6 in sequences with 10 elements, with wildcard elements specified by the 'x' character. For ease of interpretation, all features have been specified as contiguous strings of identical elements. The list includes two pairs of features that overlap with shared elements and one pair of overhanging features, in which one member is completely contained within the other. This feature set was used to generate a set of 500 sequences classified as 'functional' by overlaying the features on randomly chosen sequences until the number of features added reached a specified fraction of the size of the sequence set for each feature. Any remaining empty sequence positions were then filled in with random elements from a set of 20 possible elements. A second, completely random set of 500 sequences was also generated to represent a null, 'nonfunctional' class.

Table 2 lists the results from an application of a simple statistical feature detector to the combined set of 1000 synthetic sequences. This detector computes the chi-square metric of association between feature occurrence in a sequence and the functional class of the sequence. The table lists the top 20 features as ranked by this metric. Three types of features appear in Table 2: the synthetic 'generator' features used to produce the data, child artifacts and cross-feature artifacts. Child artifacts are composed of subsets of the elements of individual generator features, whereas cross-feature artifacts combine elements from two or more generator features.

Table 2 contains four of the generator features, ranked {4, 9, 10, 11}. The top three features in the table are cross-feature artifacts formed by the shared overlap of the generator features $\mathbf{C}_{4\rightarrow 6}$ and $\mathbf{C}_{5\rightarrow 7}$. Cross-feature overlap artifacts are particularly prominent in feature detector output, since they share the abundances of multiple generator features. Features $\{14 \rightarrow 18\}$ are non-overlap cross-feature artifacts formed from the components of the generator features $\mathbf{A}_1$, $\mathbf{A}_{10}$, $\mathbf{C}_{4\rightarrow 6}$ and $\mathbf{C}_{5\rightarrow 7}$. Non-overlap cross-feature artifacts have lower abundances than their component generator features, but they are common when multiple strong generator features are present. Features {5, 6, 7, 8, 12, 13, 19, 20} are child artifacts formed from components of individual generator features. A third type of artifact that is not shown in Table 2 is the random artifact, which contains one or more elements due to chance overrepresentation.

The results shown in Table 2 are typical of feature detector output when the data contains strong higher order features. The strongest and/or simplest generator features are ranked near the top, embedded in a shower of artifacts which displaces the other generator features down the list. The challenge of sequence feature detection is to separate the generator features from the artifacts.

No existing algorithm is specifically designed to screen artifacts of the types seen in Table 2 from sequence data, and as far as the author is aware, no algorithm has demonstrated success at extracting generator features and suppressing artifacts in a synthetic dataset that contains a complex group of features such as that shown in Table 1. Nevertheless, it is worthwhile to measure the effectiveness of existing algorithms at this task, particularly since a number of feature selection techniques are structured to reduce redundancies in the feature set. Section 1 in Supplementary Material contains a detailed evaluation of the sequence feature detection capabilities of seven existing algorithms, including a Markov model, the chi-square, ReliefF, FCBF, and mRMR filters, the SVM-RFE embedded search algorithm and a neural network wrapper, using five synthetic datasets that contain sets of generator features with varying levels of overlap and differing relative feature abundances. Results reveal that even the best algorithms all select at least 50% false positives among the top ranked features in each dataset and generally detect only the simplest and/or strongest features. In Section 4.7 in Supplementary Material, the four algorithms with the best performances on the synthetic data are applied to the laboratory HIV PR dataset, and found to be ineffective at accurately ranking features by statistical significance as measured by the methods presented in this article. Existing feature selection algorithms are therefore poorly suited to act as general detectors of localized features in sequence sets for the purpose of analysis. It should be noted that the poor performances of these algorithms at the separation of generator features from artifacts in sequence sets is not inconsistent with the strong performances which some have demonstrated at the prediction of functional class from sequence, since a set of features that includes artifacts may contain much the same information as the set of generator features, though this information is expressed in a representation that is not optimal for analysis.

Table 1. Synthetic features, with quantities added to the 'functional' sequence set listed as fractions of the size of the set

Index	Feature	Fraction	Index	Feature	Fraction
1	**A**xxxxxxxxx	0.25	6	xxxx**CCC**xxx	0.20
2	xxxxxxxxx**A**	0.25	7	**DDDD**xxxxxx	0.05
3	x**BB**xxxxxxx	0.10	8	**DDDDD**xxxxx	0.05
4	xxxxxxx**BB**x	0.10	9	**EEEEEE**xxxx	0.05
5	xxx**CCC**xxxx	0.20	10	xxxx**EEEEEE**	0.05

Table 2. Top 20 features in the synthetic data, ranked by the chi-square metric and shown with % abundances as measured in the functional set

Rank	Feature	χ^2	%Ab	Rank	Feature	χ^2	%Ab
1	xxxxCCxxxx	212	0.35	11	**A**xxxxxxxxx	87	0.29
2	xxxxCxxxxx	178	0.39	12	xxxxxxCxxx	75	0.24
3	xxxxxCxxxx	154	0.38	13	xxxCxxxxxx	73	0.24
4	xxx**CCC**xxxx	121	0.21	14	AxxxCCxxxx	63	0.12
5	xxxxCxCxxx	119	0.21	15	AxxxCxxxxx	60	0.13
6	xxxxxCCxxx	119	0.21	16	AxxxxCxxxx	60	0.13
7	xxxCxCxxxx	118	0.22	17	xxxxCxxxxA	54	0.10
8	xxxCCxxxxx	118	0.22	18	xxxxxCxxxA	53	0.10
9	xxxx**CCC**xxx	117	0.21	19	DDxDxxxxxx	52	0.10
10	xxxxxxxxx**A**	116	0.28	20	DxDDxxxxxx	52	0.10

Generator features are bolded.

2.2 Feature tables

Since a sequence set is a cross-classified, multivariate dataset in which each sequence position represents a discrete variable and each possible element at a position represents a category, it is natural to apply the framework of discrete multivariate analysis to the problem of sequence feature detection. CD therefore approaches the feature detection problem by organizing sequence data into contingency tables, as in Hu *et al.* (1993) (for which the present author developed the statistical methodology). A set of N different sequences of length l with e possible elements at each position may

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

be organized into an l-dimensional table, with each dimension divided into e categories, one for each element. Each of the e^l cells in this fundamental table represents a single possible sequence. Summary contingency tables, called configurations, may then be formed by collapsing the fundamental table across one or more dimensions (sequence positions) and accumulating the counts in the remaining dimensions. These configurations are named and indexed by the remaining dimensions, so that collapsing the four-dimensional fundamental table $\mathbf{T}_{1234}(i,j,k,l)$ across the dimension corresponding to the third sequence position, for example, produces the configuration table $\mathbf{C}_{124}(i,j,l)$. The cells of a configuration contain the abundances of all sequence features that specify elements at the positions that correspond to the dimensions of the configuration. For example, the abundance of the feature $\mathbf{A}_1\mathbf{B}_2\mathbf{C}_4$ is listed in cell $\mathbf{C}_{124}(1,2,3)$, if the elements are indexed in alphabetical order.

CD derives separate feature tables from a configuration by condensing each of its dimensions down to just two categories, which represent the presence or absence of a single feature element at each position. An e^r configuration that contains all features composed of e possible elements occurring at a particular set of r positions is separately condensed into e^r individual 2^r feature tables that are used to evaluate individual features. For example, the 2^3 feature table used to evaluate $\mathbf{A}_1\mathbf{B}_2\mathbf{C}_4$ has categories $\mathbf{A}$ and $\sim\mathbf{A}$ in dimension 1, $\mathbf{B}$ and $\sim\mathbf{B}$ in dimension 2, and $\mathbf{C}$ and $\sim\mathbf{C}$ in dimension 3. The cell in a feature table that corresponds to the presence of the feature's elements at all sequence positions contains the observed feature count and is termed the feature cell.

CD evaluates the significance of a feature by computing its P-value (Pval) given a null model that represents the background data. The evaluation of a first-order feature is straightforward. The table for feature $\mathbf{A}_1$, for example, has just one dimension with two cells, labeled $\mathbf{A}_1$ and $\sim\mathbf{A}_1$. In this case, an appropriate null model might specify equal probabilities of $(1/e)$ for finding any member of a set of e elements at the sequence position. The expected count for any element would then be computed as $(1/e)\times N$. An alternative null model substitutes abundance fractions measured from a larger, background dataset, such as amino acid fractions in the proteome, for the element probabilities. In either case, a Pval is obtained for the feature table by summing the exact binomial probabilities of all tables with the same margin (same N) that are at least as improbable as the observed table.

2.3 Log-linear models for feature tables

Background models for higher order feature tables must specify the underlying structure in the table. Log-linear methods are widely used to model contingency tables (Bishop *et al.*, 1975). By analogy with analysis of variance methods, log-linear models represent the logarithms of the expected cell counts in a table as sums of sets of effects with orders up to the order of the table. For example, the log expected count for cell (i,j) in the complete, or 'saturated' log-linear model for a 2D table with I rows and J columns is:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{1}$$

where:

$$u = \frac{1}{IJ}\sum_{i,j}\log m_{ij} \tag{2}$$

is the mean effect which applies to all cells,

$$u_{1(i)} = \frac{1}{J}\sum_{j}\log m_{ij} - u \tag{3}$$

is the effect of the variable associated with the first table dimension,

$$u_{2(j)} = \frac{1}{I}\sum_{i}\log m_{ij} - u \tag{4}$$

is the effect of the variable associated with the second dimension, and:

$$u_{12(ij)} = \log m_{ij} - (u + u_{1(i)} + u_{2(j)}) \tag{5}$$

is the interaction effect between variables 1 and 2. With the exception of the single-valued mean effect, each effect, or u-term, contains as many indexed parameters as there are cells in the configuration that shares its dimensional labels. Each indexed u-term parameter contributes to the log of the counts for the subgroup of cells in the table that shares the parameter's indices. As Equations (3), (4) and (5) demonstrate, the value of each u-term parameter is defined as the difference between the mean (or single value for the top-level effect) of the logs of the counts in the cells covered by that parameter and the sum of all parameters from related lower order u-terms that cover the same cells. Each u-term therefore represents a deviation from the model composed of all related underlying effects. Equivalently, each u-term of order r can be expressed as a deviation of a u-term of order $(r-1)$ across the categories of an additional dimension. For example, u_{12} can be expressed as the deviation of u_1 across the categories of variable 2, and this two-factor effect measures the difference in the proportional distribution in one dimension between the categories of the other. A log-linear model is classified as hierarchical if every higher order u-term included in the model is accompanied by all of its lower order relatives, which are the lower order u-terms that share any dimension labels with the higher order term.

Log-linear modeling is commonly used in a procedure designed to identify the simplest set of effects capable of producing expected counts that closely match the observed table counts. In this procedure, a range of incomplete, or unsaturated hierarchical models, the selection of which may be guided by the magnitude of the u-terms in the saturated model or by outside information, is tested against the data. Since the summary configuration table that shares a u-term's dimensional labels contains all the information required to produce a set of unique maximum likelihood estimates (MLEs) of cell counts consistent with the u-term and all of its lower order relatives (Birch, 1963), MLEs can be generated for any hierarchical model directly from the configurations using iterative proportional fitting. This procedure begins by assigning uniform counts to all cells and proceeds to adjust these counts to be consistent with the configurations that correspond to the u-terms present in a model over a series of passes. The observed counts are then compared against the MLEs to measure the fit of each model. The u-terms that significantly improve the fit (by increasing the Pval) are considered likely to represent important effects and are retained in the model, while those that do not are discarded. It should be noted that since the MLEs are generated directly from the marginal proportions (the configurations) without the use of any log-linear techniques, this method may be thought of simply as proportional modeling.

The goal of table modeling is usually to find the simplest model which can reproduce the counts for all cells in a table with reasonable accuracy, and to deduce, from the structure of this model, which data dimensions or combinations of dimensions are important. To the extent that single cells are considered, they are generally evaluated as outliers from a model judged optimal for an entire table. By contrast, the focus of feature detection is on the individual combination of elements, which constitutes a feature and corresponds to the feature cell in a feature table. The goal here is the evaluation of the significance of the count in the feature cell against a background model for the cell which represents the data in the absence of the feature.

CD evaluates a feature of interest (FOI) by constructing a background model which includes the u-terms of all orders associated with any child features that are significant in the absence of the FOI, but excludes u-terms that are associated only with the feature itself. The model used to evaluate second-order features is simple. Since a second-order feature introduces a tendency for two elements to occur together, it generates a difference in the proportional distribution in each dimension across the categories of the other, which corresponds to the second-order term u_{12}. This term is therefore excluded from the background model. Both first-order effects must be included in the model, however, to account for the inherent numerical asymmetry between the two categories in each dimension which occurs because one category represents the presence of a particular element, while the other represents the presence of all other elements. The background model therefore includes the terms u, u_1 and u_2. Before a final background MLE

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

representing the absence of the FOI can be generated, the effects of any overrepresentation of the FOI on the configurations in the model must be removed. This is done by applying an iterative smoothing process in which single counts are removed from the FOI's feature cell in each step. After each step, the configurations are recomputed and the MLE is recalculated. The process continues until the count in the feature cell is no longer significantly overrepresented compared with its MLE. The resulting smoothed MLE, produced by the removal of R sequences from the feature cell, represents the expected count of the feature in a dataset of size $(N-R)$ in which the feature is absent as an effect, since its count merely conforms to the underlying model. To generate the final background MLE, this smoothed MLE is multiplied by a factor of $N/(N-R)$. This computation applies the mean proportion of the unsmoothed table of size N to the smoothed background model, scaling the MLE up so that it is appropriate for the evaluation of the feature count in the unsmoothed table. A statistical metric is then computed to measure the distance between the observed feature count and the scaled background MLE, and compared against the chi-square distribution with one degree of freedom (which represents the variation of the feature) to generate the feature's Pval.

2.4 *u*-term cascades

While a second-order feature is associated with a single u-term (u_{12}) not included in the background model, higher order features present a more complex picture. The third-order feature $\mathbf{A}_1\mathbf{B}_2\mathbf{C}_3$, for example, contains three child features composed of pairs of elements that occur together. Each child feature introduces a two-factor effect into $\mathbf{A}_1\mathbf{B}_2\mathbf{C}_3$'s feature table: $\mathbf{A}_1\mathbf{B}_2$ introduces u_{12}, $\mathbf{A}_1\mathbf{C}_3$ introduces u_{13} and $\mathbf{B}_2\mathbf{C}_3$ introduces u_{23}. $\mathbf{A}_1\mathbf{B}_2\mathbf{C}_3$ itself introduces a three-factor effect, u_{123}, which reflects the simultaneous tendencies for $\mathbf{A}_1\mathbf{B}_2$ to occur in category $\mathbf{C}$ but not $\sim\mathbf{C}$ of position 3, $\mathbf{A}_1\mathbf{C}_3$ to occur in category $\mathbf{B}$ but not $\sim\mathbf{B}$ of position 2 and $\mathbf{B}_2\mathbf{C}_3$ to occur in category $\mathbf{A}$ but not $\sim\mathbf{A}$ of position 1. An extension of this analysis to higher orders reveals that the synergy associated with a feature of order r is represented in a log-linear model by a cascade of interaction effects of orders r and below. Each higher order effect in the cascade represents the deviation of the effects immediately beneath it across an additional dimension.

Table 3 displays the complete cascade of interaction effects present in the feature table for the synthetic fourth-order feature **ABCD**. This feature was introduced into 20% of 1000 otherwise randomly generated sequences, each constructed from a set of 20 possible elements. The table lists the parent feature and its third- and second-order child features, along with the u-term parameters from the saturated log-linear model which represent the highest order effects associated with each feature. Two measures of u-term strength are listed: the standardized u-term magnitude and its Pval, which measures the probability that a u-term of a given magnitude or larger would occur by chance. Table 3 demonstrates that **ABCD** is modeled with a cascade of u-terms, which generally increase in strength with decreasing order. This is the characteristic signature of an overrepresented higher order feature in data that is otherwise random.

The configuration tables formed by the collapse of a feature table and the accumulation of its cell counts into the remaining dimensions form feature tables for the child features. For example, summation over the fourth dimension of the 2^4 feature table used to evaluate **ABCD** produces the 2^3 feature table for its child feature, **ABC**x. Furthermore, since a configuration completely specifies its associated u-term and all related lower order u-terms, each configuration of a feature table specifies the complete u-term cascade for a child feature.

Table 3. A fourth-order u-term cascade with associated child features

Feature	Effect	Mag	Pval	Feature	Effect	Mag	Pval
ABCD	u_{1234}	1.63	0.103	**AB**xx	u_{12}	3.37	0.001
				Ax**C**x	u_{13}	3.05	0.002
ABCx	u_{123}	1.69	0.092	**A**xx**D**	u_{14}	2.80	0.005
ABx**D**	u_{124}	2.56	0.010	x**BC**x	u_{23}	2.83	0.005
Ax**CD**	u_{134}	2.25	0.024	x**B**x**D**	u_{24}	2.99	0.003
x**BCD**	u_{234}	2.47	0.013	xx**CD**	u_{34}	3.30	0.001

The highest order u-term associated with each feature is listed, along with its normalized magnitude and Pval.

2.5 Cascade detection

While it is immediately clear which u-terms must be included in the background model for a second-order FOI (u, u_1 and u_2), the construction of the background model for a higher order feature is more involved. In order to account for the effects which any significant background features may have on the abundance of the FOI, CD must detect any underlying child features in its feature table which are significantly overrepresented in the absence of the FOI, and include their associated configurations in the background model. Since underlying child features must be evaluated before an FOI can be evaluated, and the evaluation of the child features themselves requires the evaluation of any grandchild features and so on downwards, CD must implement a descending, recursive procedure which constructs and tests background models for all underlying features before evaluating the FOI.

The set of u-terms excluded from the FOI's background model, together with the contributions which the FOI makes to the u-terms included in the model, are termed the feature cascade. The comparison of the observed count in the feature cell with its background MLE evaluates the contribution of the feature cascade to the feature count, enabling the detection of cascades that make significant contributions. These cascades represent the packets of synergy not present in the background data that boost feature counts above their MLEs.

A detailed presentation, with flow charts, of the complete CD algorithm is given in Section 2 in Supplementary Material. The presentation includes descriptions of CD's method of screening out child artifacts by iteratively smoothing parent features, CD's technique of detecting variable-position features by associating features that differ only by position shifts into groups, CD's application of the Simes procedure (Benjamini and Hochberg, 1995) to control false positives and other relevant information about the algorithm.

2.6 Synthetic results

Table 4 shows the results of an application of CD to the set of 500 ‘functional’ synthetic sequences which was produced using the set of generator features with overlaps listed in Table 1, and analyzed with the chi-square metric in Table 2. Table 4 displays all features detected by the algorithm using a false positive fraction of 0.01 and an abundance threshold of 2%, arranged by feature order for clarity and shown with negative $\log_{10}P$vals (pPvals, higher is more significant). Also listed with each feature are the underlying child features which CD detects and includes in the background models for each feature table (except for first-order children, which are always included), and the degrees of freedom (DOF) fixed in each background model compared with the total DOF in the feature table (which is 2^r for a feature of order r). Each child feature introduces its own u-term cascade into the background model, and each u-term in a cascade fixes a single DOF.

Table 4. CD results from the synthetic dataset generated by the features in Table 1, arranged by feature order and shown with significant child features, degrees of freedom fixed in each model compared with total DOF (FDF column), and negative $\log_{10}P$vals (pPval column)

Rank	Feature	Child	FDF	pPval	Rank	Feature	Child	FDF	pPval
1	$\mathbf{A}_1$	–	1/2	67	4	$\mathbf{C}_{5\to7}$	$\mathbf{C}_5\mathbf{C}_6$	5/8	53
2	$\mathbf{A}_{10}$	–	1/2	65	10	$\mathbf{D}_{1\to4}$	–	5/16	16
5	$\mathbf{B}_2\mathbf{B}_3$	–	3/4	35	9	$\mathbf{D}_{1\to5}$	$\mathbf{D}_{1\to4}$	17/32	18
6	$\mathbf{B}_8\mathbf{B}_9$	–	3/4	31	7	$\mathbf{E}_{1\to6}$	$\mathbf{E}_5\mathbf{E}_6$	8/64	19
3	$\mathbf{C}_{4\to6}$	$\mathbf{C}_5\mathbf{C}_6$	5/8	57	8	$\mathbf{E}_{5\to10}$	$\mathbf{E}_5\mathbf{E}_6$	8/64	19

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

The results in Table 4 contain only the 10 generator features from Table 1, demonstrating that CD recovers all synthetic features while suppressing all the artifacts which they produce among the set of 200 first-order and 364 higher order features that meet the specified 2% abundance threshold. The table shows that any overlaps between other generator features and an FOI are included in the background model for the FOI. For example, the models for the features $\mathbf{C}_{4\to6}$ and $\mathbf{C}_{5\to7}$ each include the child feature $\mathbf{C}_5\mathbf{C}_6$ that represents their mutual overlap. These background models automatically fix one DOF for the mean effect and one DOF for each of the three first-order effects, whereas the $\mathbf{C}_5\mathbf{C}_6$ overlap child feature fixes a second-order effect, so that five DOF out of a total of eight are fixed in each model. In a similar fashion, the $\mathbf{E}_5\mathbf{E}_6$ overlap child feature is fixed in the models for $\mathbf{E}_{1\to6}$ and $\mathbf{E}_{5\to10}$. Since the feature $\mathbf{D}_{1\to5}$ overhangs the feature $\mathbf{D}_{1\to4}$, the model for $\mathbf{D}_{1\to5}$ contains $\mathbf{D}_{1\to4}$ as a child feature. Six DOF are fixed automatically, and $\mathbf{D}_{1\to4}$ fixes one fourth-order effect, 4 third-order effects and 6 second-order effects, so that 17 of 32 DOF are fixed for $\mathbf{D}_{1\to5}$.

The higher order generator features are smoothed during the evaluation of each of their numerous child artifacts, rendering these artifacts insignificant and excluding them from the results. The feature $\mathbf{D}_{1\to5}$ is smoothed during the evaluation of $\mathbf{D}_{1\to4}$, since it is a significant parent, but $\mathbf{D}_{1\to4}$ remains significant.

CD produces results similar to those of Table 4 for all five synthetic datasets, as shown in Section 3 in Supplementary Material. For all datasets, the algorithm ranks all generator features above all artifacts. CD includes one false positive, a random artifact, among the 51 total features which it detects in all datasets, yielding an overall false positive fraction of 0.0196, not inconsistent with the specified estimated value of 0.01. As a further check, CD was applied to a pure 'noise' dataset of 10 000 sequences that contained no generator features. After evaluating the 8 806 features of second-order and higher present in this dataset with abundances above threshold, the algorithm detected no features that satisfied the 0.01 false positive fraction cutoff. The average computed *P*val for all of these features was 0.49, which is consistent with the expected value of 0.5 for noise.

3 RESULTS

3.1 HIV PR cleavage specificity

HIV PR is a homodimeric, aspartic protease that cleaves the gag and gag-pol viral polyproteins during virus assembly and is required for HIV viability (Kohl *et al.*, 1988). HIV PR recognizes an amino acid sequence of eight residues centered on the cleavage site, which is designated P_4-P_3-P_2-P_1↓ P'_1-P'_2-P'_3-P'_4 (Schechter and Berger, 1967), with cleavage at ↓ between positions P_1 and P'_1. The mechanisms by which the enzyme recognizes its substrate are not well understood due to a lack of obvious features shared by the cleaved peptides. HIV PR inhibitors constitute a major component of current HIV therapy, but the rapid mutation rate of HIV threatens the effectiveness of these treatments (Lu, 2008). The development of a more complete understanding of HIV PR specificity is therefore important. Techniques useful in this effort should also prove valuable in the elucidation of the specificities of the proteases of other potent viruses, such as hepatitis C, and of proteases in general, including >500 varieties in humans. For a brief review of machine learning techniques applied to this problem, see You *et al.* (2005). For a recent study, see Rögnvaldsson *et al.* (2009).

CD was applied to the HIV PR cleavage dataset of Schilling and Overall (2008), which was generated using proteome-derived peptide libraries as part of a larger effort designed to test PICS (Proteomic identification of protease cleavage sites), the authors' high-throughput analysis technique. This dataset was chosen because it was generated from a proteome scan, and is therefore less likely to contain the strong sampling bias (and the resultant spurious features of both first order and higher order) that are present in summary datasets which include cleavage sites generated by limited mutagenesis of a small set of cleaved exemplars. The PICS authors conducted a pilot analysis of their data, including a specificity matrix and a first look at a few potential pair features, but no statistical tests were performed. The PICS data is composed of two sets of cleavage sites generated by the application of HIV PR to two tryptic peptide libraries derived from HEK cell lysis. The set of 211 cleaved octamers extracted here from the dataset listed in Supplementary Table 20 of the PICS paper has 63% of its sequences in common with the set of 343 octamers extracted from the dataset listed in Supplementary Table 19 of the paper. Since analysis by CD revealed that the two sets of octamers have very similar first-order feature structures (as can also be seen in the heat maps in Figure 6 of the PICS study), and results also showed similarity in the higher order structure, the two sequence sets were combined (eliminating duplicates) to form a single set of 412 cleaved octamers.

	P_4	P_3	P_2	P_1	↓	P_1'	P_2'	P_3'	P_4'	P_4	P_3	P_2	P_1	↓	P_1'	P_2'	P_3'	P_4'
D	**4**		**12**	**3**	↓	*3*	*3*		**5**					↓	*3*			
E	**6**	**6**	**52**		↓		**42**			**2**		**6**	**99**	↓	*5*		*4*	*5*
Q			*4*	*2*	↓	*5*							*5*	↓			**3**	
N					↓	*4*							*7*	↓		*2*		
H			*3*	*3*	↓		*4*	**5**				**2**	*3*	↓				
P	**3**	*3*	*8*	*9*	↓	X	*7*	*7*					*8*	↓	X	*4*	**8**	**3**
Y			*3*	**7**	↓		*4*		*3*				*5*	↓		**2**		
W			*2*		↓									↓				
S			*5*	*2*	↓	*3*	*2*						11	↓				**2**
T				*9*	↓								*8*	↓				
G			*5*		↓	*8*	*8*						*5*	↓	**5**			
A			**2**		↓				**2**				*8*	↓	**7**	**6**	**3**	
M				**7**	↓	**3**							3	↓				
C					↓	*2*							3	↓				
F		**4**	*4*	**12**	↓	**8**	*2*	**4**					*5*	↓				
L		**7**	*7*	**24**	↓	**16**		**4**					*12*	↓		**3**		
V			**8**	*9*	↓	**8**	**10**						*11*	↓				
I			**4**	*8*	↓	**8**	**5**						*8*	↓	**2**			

Fig. 1. Left pane: negative $\log_{10}$*P*vals (higher is more significant) for first-order HIV PR cleavage features from the PICS dataset. Right pane: first-order GluC features from PICS. Higher order features were not smoothed, so the *P*vals reflect the contributions of amino acids to features of all orders.

The left pane of Figure 1 displays the exact binomial *pP*vals for the first-order HIV PR features formed from particular amino acids, computed using the overall frequency of each amino acid in the human proteome as the estimate of the probability of finding that amino acid at a position in a sequence. Overrepresented (favorable) features are bolded, while underrepresented (unfavorable) features are shown in italics. All features meeting a *P*val threshold of 0.01 are displayed in order to reveal patterns. Rows are arranged in top-to-bottom order of increasing amino acid hydropathy. The basic amino acids **R** and **K** are excluded, because the tryptic peptide libraries lack internal basic residues, and the feature ↓$\mathbf{P}_1$ (proline at P'_1) is omitted because cleavage at proline yields secondary amines, which are less reactive to the esters used by PICS to tag the peptide fragments, and may effect the results.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

The feature set in Figure 1 contains strong negatively charged and hydrophobic features, along with patterns which classify the hydrophobic features by side chain geometry. The features also show substantial symmetry about the cleavage site. For a detailed analysis with structural insights, see Section 4.1 in Supplementary Material.

Analysis of the higher-order features in the PICS data that are formed from particular amino acids suggests that favorable cooperativity is mostly distributed across a set of relatively weak features which each play small roles in the determination of cleavage. When a 10% false positive criterion for statistical significance is applied, just two features are detected: $\mathbf{L}_1{\downarrow}\mathbf{N}_2$ (*P*val = 1.8E-5, abundance 2.7%) and $\mathbf{F}_3\mathbf{G}_1{\downarrow}$ (*P*val = 2.9E-5, abundance 2.7%). The more liberal 50% false positive criterion, for which half of the detected features are considered statistically significant, yields a *P*val cutoff of 0.039. At this lower significance level, 54 features are detected with typical abundances of 1–4% of the sequence set. The feature set includes multiple examples that combine two or more of the strong first-order hydrophobic or charged features shown in Figure 1. The detection of the charged triplets $\mathbf{E}_3{\downarrow}\mathbf{E}_1\mathbf{E}_2$ (*P*val = 0.002, abundance 1.2%) and $\mathbf{E}_2\mathbf{E}_1{\downarrow}\mathbf{E}_2$ (*P*val = 0.027, abundance 1.7%) suggests favorable cooperativities between glutamate residues in the neighborhoods of the two most significant first-order features, $\mathbf{E}_2{\downarrow}$ and $\mathbf{E}_2{\downarrow}$. For a detailed analysis, see Section 4.2 in Supplementary Material.

The significance of multiple first-order hydrophobic and negatively charged features prompted a general investigation of higher order features that combine hydrophobic and/or negatively charged residues. When the hydrophobic and negatively charged amino acids are each replaced in the data with single symbols (shown here as **H** and **N**), the feature $\mathbf{N}_2{\downarrow}\mathbf{H}_1\mathbf{H}_2\mathbf{H}_3$(*P*val = 0.004, abundance 12.6%) is detected at the 50% false positive level, suggesting possible cooperativity between negatively charged residues at position P_2 and a concentration of hydrophobicity on the prime side formed from first-order hydrophobic features detected in Figure 1. Results also identify several other overrepresented features that combine negative charge with hydrophobicity, and show a set of prime-side hydrophobic pairs $\{{\downarrow}\mathbf{H}_1\mathbf{H}_2,\ {\downarrow}\mathbf{H}_1\mathbf{H}_3,\ {\downarrow}\mathbf{H}_1\mathbf{H}_4\}$ with weak but consistent overrepresentation. Notably absent from the results, however, is evidence of cooperativity associated with the feature $\mathbf{H}_1{\downarrow}\mathbf{H}_1$, which represents the simultaneous presence of hydrophobic residues on each side of the cleavage site and has been used to classify cleaved substrates (Beck *et al.*, 2002). For a detailed analysis of hydrophobic/negatively charged features, see Section 4.3 in Supplementary Material.

CD can address the question of whether the PICS data contain evidence for synergy associated with any features favorable to cleavage that have been highlighted by previous studies. For example, lists of the 10 most important favorable amino acid pairs are presented in You *et al.* (2005). These pairs all combine elements from the lists of important single-position features presented in the study, suggesting that they may represent cross-feature artifacts. CD detects no significant synergy associated with any of these pairs in the PICS data; all the pairs have abundances close to their expected counts. The PICS authors suggest the possibility of favorable cooperativity between $\mathbf{L}_3{\downarrow}$ and $\mathbf{A}_1{\downarrow}$. The feature $\mathbf{L}_3\mathbf{A}_1{\downarrow}$ is overrepresented in CD results, with a *P*val of 0.008 and an abundance of 3.2%. Strong unfavorable pair cooperativities, probably associated with steric interference between larger amino acid side chains in the substrate, have been detected using biochemical background modeling (Ridky *et al.*, 1996). Although CD does not evaluate unfavorable cooperativity in datasets classified as functional (Section 2.4 in Supplementary Material), results do reveal that all but one of these pair features has zero abundance in the data, which is consistent with an unfavorable effect.

The right pane of Figure 1 shows first-order results derived from a set of 371 sequences extracted from the PICS data for GluC, a protease with a well-defined, ‘canonical’ cleavage feature ($\mathbf{E}_1{\downarrow}$). As might be expected, GluC’s feature set outside of the canonical site is limited compared with that of HIV PR, at both first-order and higher order. GluC shows first-order preferences for small side chains (**G** and **A**) in the prime-side neighborhood of the cleavage site. The liberal 50% false positive threshold results in the detection of just one higher order feature with an abundance of >2%: $\mathbf{E}_3\mathbf{E}_2{\downarrow}$, with a *P*val of 0.0006 and an abundance of 3.5%. This feature suggests that an additional concentration of negative charge adjacent to the canonical feature $\mathbf{E}_1{\downarrow}$ favors cleavage.

3.2 Structure–function results for the Schellman loop

The Schellman loop (Schellman, 1980) is a six-residue, C-terminal helix-capping motif in proteins. In helix nomenclature, the motif covers positions C3-C2-C1-CCap-C'-C'' (Aurora and Rose, 1998). The motif is defined by the presence of two hydrogen bonds, which link backbone carbonyl oxygens at C3 and C2 with the backbone nitrogens at C'' and C', respectively. Glycine is very often found at the C' position. To the extent that other particular amino acids are structurally important, they are thought to contribute to stability mostly through hydrophobic/hydrophilic effects and possible hydrogen bonding with the backbone. As far as the author is aware, no general study of cooperativity in the Schellman loop has been performed. Munoz (1995) described pairs of interacting hydrophobic residues in N-terminal helix caps, and such a ‘hydrophobic staple’ was identified by Aurora and Rose (1998) at the (C3, C'') positions in the Schellman loop. The latter study also reported that attractive electrostatic interactions (salt bridges) sometimes take the place of hydrophobic staples in helix caps.

CD was applied here to a set of 4682 Schellman loop sequences extracted from the PDB using PDBeMotif (Golovin and Henrik, 2008). First-order results, analyzed in Section 4.4 in Supplementary Material, include highly significant features containing glycine, alanine and charged, polar and hydrophobic residues. Higher order results also contain many statistically significant features. A false positive fraction criteria of 10%, which yields a *P*val cutoff of 0.005, identifies 106 significant features of second- through fourth-orders. However, since there is prior justification for focusing interest on some particular features, including possible hydrophobic or electrostatic interactions, a more liberal absolute *P*val cutoff of 0.01 was applied, resulting in the identification of the 99 second-order, 36 third-order and 6 fourth-order features listed in Section 4.5 in Supplementary Table 10. The higher order feature with the lowest *P*val is xxx**AG**x, with an abundance of 11.9% and a *P*val of 7.5E-22. Other features that combine **GC**’ with an alanine residue elsewhere in the loop are also present in the data, including xx**A**x**G**x and xxxx**GA**, with abundances of 10.8 and 6.5%, respectively, and xxx**AGA** and xx**EAGA**. Alanine is also present paired with itself in three features: xxx**A**x**A**, **A**x**A**xxx and xx**AA**xx. While the first-order favorability of alanine at positions C3 → CCap (Section 4.4

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 5. Schellman loop pair features that combine one of the positively charged amino acids Lys(**K**) or Arg(**R**) with a negatively charged Asp(**D**), Glu(**E**), or the similar, strongly polar Gln(**Q**) or Asn(**N**)

Feature	*P*val	%Ab	Obs:Exp	Feature	*P*val	%Ab	Obs:Exp
xxx**R**x**E**	1.2E-12	1.1	49:12	xx**EK**xx	1.2E-6	1.9	87:48
xxx**E**x**K**	3.0E-9	1.1	50:17	xx**RD**xx	7.7E-3	0.45	21:11
xxx**E**x**R**	3.1E-5	0.68	32:13	xx**KE**xx	9.0E-3	1.2	54:37
xxx**D**x**R**	2.5E-4	0.38	18:6				
xxx**K**x**E**	7.0E-3	0.62	29:16	xxxx**RD**	2.5E-3	0.32	15:6
				x**KE**xxx	3.3E-3	2.4	110:82
Kxxxx**E**	1.6E-6	0.47	22:5	xxx**NK**x	2.1E-4	0.81	38:19
Rxxxx**E**	2.1E-6	0.51	24:6	**N**xxx**K**x	2.8E-3	0.38	18:8
Exxxx**R**	5.7E-5	0.38	18:5	**Q**xxxx**K**	2.9E-3	0.32	15:6
Dxxxx**K**	6.7E-5	0.49	23:8	xxx**Q**x**R**	4.1E-3	0.49	23:11
Dxxxx**R**	5.2E-4	0.36	17:6	**N**xxxx**K**	7.3E-3	0.34	16:7
Kxxxx**D**	6.8E-3	0.26	12:5	**R**xxxx**Q**	9.5E-3	0.26	12:5

Charged pairs are grouped by position, while pairs with polar members are grouped at the end of the list. Features are listed with *P*val, % abundance, and observed and expected counts. Higher order feature smoothing was turned off, so that the *P*vals reflect the contributions of electrostatic pairs to cooperativities of all orders (xx**E**x**R**, xx**KE**xx and xx**EK**xx contribute to higher order features).

in Supplementary Material) suggests that its presence at individual positions favors loop formation as it favors helix formation, by minimizing steric hindrances without providing a helix-breaking degree of flexibility, the higher order findings suggest the presence of cooperativities between alanine and **GC**′ or other alanines in the motif.

Higher order Schellman loop results also show multiple, statistically significant features composed of pairs of amino acids with oppositely charged side chains, suggesting that attractive electrostatic interactions play a role in stabilizing the motif. Table 5 displays the 16 oppositely charged pair features that meet the *P*val threshold of 0.01, along with 6 features that combine a positively charged amino acid with the strongly polar **Q** or **N**, which are structurally similar to the negatively charged **D** or **E**.

Table 5 shows that multiple oppositely charged pairs occur at both (C3, C″) and (CCap, C″), which are the positions of the two hydrophobic staples in the motif (see below), and also at (C1, CCap). Structural information from Motivated Proteins (Leader and Milner-White, 2009) suggests that side chains at (C1, CCap) can plausibly interact, as they can at the staple positions. Six of the possible eight oppositely charged pairs that can form from {**K**, **R**, **D**, **E**} at (C3, C″) occur as features in Table 5. In addition to the electrostatic pairs shown in Table 5, the higher order results also contain 25 other pair features that combine charged and polar or polar and polar amino acids, including 12 that occur at (C3, C″), (CCap, C″) or (C1, CCap). While the abundances of individual electrostatic pairs are small, 13.2% of Schellman loop sequences contain one or more of the pairs listed in Table 5. Just two pair features which meet the *P*val threshold of 0.01 contain like charges: **E**xx**D**xx and xxxx**KK**.

The higher order Schellman loop results also contain the feature xx**RE**x**R** (*P*val = 0.0005), which includes opposite-charge pairs at both (C1, CCap) and (CCap, C″), and may represent an attractive triplet interaction. The feature x**K**x**KK**x (*P*val = 0.0017) may represent a concentration of positive charge that stabilizes the loop by interacting with local carbonyl backbone groups, helping to offset the negative charge at the C-terminal end of the helix dipole.

The significance of attractive electrostatic interactions is emphasized when the positively charged amino acids {**K**, **R**} and the negatively charged amino acids {**D**, **E**} are replaced in the sequence set with single symbols representing general positive and negative charges. When this is done, the four possible general attractive electrostatic pairs at the two hydrophobic staple locations, which are {+xxxx−, −xxxx+, xxx+x− and xxx−x+}, constitute four of the six most significant higher order features in the data, with *P*vals between 5.4E-12 and 4.5E-21. Results from this analysis also include the feature xxx−+− (*P*val = 0.0059), which may represent a general attractive triplet interaction.

The presence in the higher order Schellman loop results of multiple hydrophobic (such as **L**xxxx**V**) and amphipathic (such as x**L**xxx**E**) pair features, in addition to the charged/polar pairs described above, prompted a general investigation of higher order features that contain hydrophobic and/or polar residues. The results, presented in detail in Section 4.6 in Supplementary Material, confirm that the previously observed ‘hydrophobic staple’ at (C3, C″) represents strong cooperativity and reveal a second highly significant, abundant hydrophobic pair at (CCap, C″) that is likely to represent a second staple. Results also reveal a set of highly significant and abundant amphipathic features, suggesting that amphipathic cooperativity is a major design principle in the motif.

4 DISCUSSION

Cascade detection, a new algorithm for the extraction of localized features from sequence sets, has been presented. CD is a natural extension of the proportional modeling techniques used in contingency table analysis into the domain of the detection of multifactor synergies like sequence features. The algorithm detects the packet of synergy associated with a significant higher order feature by comparing the observed feature count with the expected count generated by a background model that represents the data in the absence of the feature. Since the background model is generated by smoothing, the observed and expected counts for the feature can be derived from a single dataset, and the model contains all significant structure that exists in the data except that which is associated exclusively with the feature.

CD has been successfully tested on synthetic data and applied to the analysis of sets of amino acid sequences from problems in two very different domains: the characterization of HIV PR cleavage specificity and the analysis of structure–function factors in the Schellman loop. Favorable cooperativity in the HIV PR dataset appears to be mostly weak and broadly distributed, while much stronger favorable synergies were detected in the Schellman loop data. These findings suggest the possibility that stronger favorable cooperativities exist in systems in which features play a structural role than in systems where features have a more transient, sensing/recognition function. More work is required, however, before any such conclusion can be drawn, and it is also important to note that the Schellman loop dataset contains >11 times as many sequences as the HIV PR dataset, and therefore has much better statistics. It is likely that, given a larger dataset of cleaved sequences, the signals of some cooperative features in the HIV PR data would rise further above the noise level (although at the same time, the signals of other features would likely weaken).

It is also worth noting that the degree to which a set of features detected in naturally occurring sequence data represents

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

the theoretically optimal set for the promotion of function depends on the thoroughness with which evolution has sampled the space of all possible features and the efficiency of the natural selection process. The most effective combinations of amino acids may not yet have been found, or may have been found in some proteins but not yet in others, so that some features which are potentially highly favorable may be absent from the data, or may not exhibit a statistical significance that matches their importance.

Potential further applications of CD include analyses of other protein motifs (including the 12 additional small motifs cataloged in PDBeMotif), analyses of the cleavage specificities of other proteases and studies of binding sites (in proteins and possibly nucleic acids), phosphorylation sites and sorting signals. The algorithm is well suited to work with high-throughput proteomic techniques or combinatorial mutagenesis methods to characterize cooperativity within a single protein or between interacting proteins. Although CD was developed for sequence analysis, the algorithm should prove useful more generally for the detection of multifactor synergies in a wide range of datasets that can be organized into multicategory contingency tables, including results from clinical studies of disease or treatment.

ACKNOWLEDGEMENTS

The author wishes to thank Athena Newell for useful discussions and help with the manuscript, Franklyn Turbak for useful discussions and Catherine Newell for support.

Funding: Funding provided by the author.

Conflict of Interest: none declared.

REFERENCES

Aurora,R. and Rose,G.D. (1998) Helix capping. *Prot. Sci.*, **7**, 21–38.

Beck,Z.Q. *et al.* (2002) Defining HIV-1 protease substrate selectivity. *Curr. Drug Targets Infect. Disord.*, **2**, 37–50.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B,* **57**, 289–300.

Birch,M.W. (1963) Maximum likelihood in three-way contingency tables. *J. R. Stat. Soc. Ser. B*, **25**, 220–233.

Bishop,Y.M.M. *et al.* (1975) *Discrete Multivariate Analysis.* MIT Press, Cambridge, Massachusetts.

Fogel,G.B. (2008) Computational intelligence approaches for pattern discovery in biological systems. *Brief. Bioinform.*, **9**, 307–316.

Golovin,A. and Henrik,K. (2008) MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics*, **9**, 312.

Guyon,I. *et al.* (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.*, **46**, 389–422.

Hu,J.C. *et al.* (1993) Probing the roles of the residues at the e and g positions of the GCN4 leucine zipper by combinatorial mutagenesis. *Prot. Sci.*, **2**, 1072–1084.

Kohl,K.E. *et al.* (1988) Active human immunodeficiency virus protease is required for viral infectivity. *Proc. Natl Acad. Sci. USA*, **85**, 4686–4690.

Leader,D.P. and Milner-White,E.J. (2009) Motivated Proteins: a web application for studying small three-dimensional protein motifs. *BMC Bioinformatics*, **10**, 60.

Leek,J.T. and Storey,J.D. (2008) A general framework for multiple testing dependence. *Proc. Natl Acad. Sci. USA*, **105**, 18718–18723.

Lu,Z. (2008) Second generation HIV protease inhibitors against resistant virus. *Expert Opin. Drug Discov.*, **3**, 775–786.

Munoz,V. *et al.* (1995) The hydrophobic-staple motif and a role for loop-residues in α -helix stability and protein folding. *Nat. Struct. Biol.,* **2**, 380–385.

Peng,H. *et al.* (2005) Minimum redundancy maximum relevance feature selection. *IEEE Intell. Syst.*, **20**, 70–71.

Ridky,T.W. *et al.* (1996) Human immunodeficiency virus, type I protease substrate specificity is limited by interactions between substrate amino acids bound in adjacent enzyme subsites. *J. Biol. Chem.*, **271**, 4709–4717.

Rögnvaldsson,T. *et al.* (2009) How to find simple and accurate rules for viral protease cleavage specificities. *BMC Bioinformatics,* **10**, 149.

Saeys,Y., *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Schechter,I. and Berger,A. (1967) On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Comun.*, **27**, 157–162.

Schellman,C. (1980) The α_L-conformation at the ends of helices. In Jaenicke,R. (ed.) *Protein Folding.* Elsevier/North-Holland, New York, pp. 53–61.

Schilling,O. and Overall,C.M. (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat. Biotechnol.*, **26**, 685–694.

You,L. *et al.* (2005) Comprehensive bioinformatics analysis of the specificity of human immunodeficiency virus type I protease, *J. Virol.*, **79**, 12477–12486.

Yu,L. and Liu,H. (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In *ICML-03*. AAAI press, pp. 856–863.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016