Appeared in A. Okada et al. (eds.) *Cooperation in Classification and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin, Heidelberg (2009) 115–122

Analyzing protein–protein interaction with variant analysis

Gunter Ritter and María Teresa Gallegos

Faculty of Computer Science and Mathematics
University of Passau
D-94030 Passau, Germany

Abstract. Biochemistry teaches that the functioning of a biological organism is to a large extent determined by the interaction of biomolecules. One of these interactions is that of two proteins. Besides experimental methods, sequence information in combination with computational methods can be used to shed light in their mode of operation. For this purpose we present a method based on variant analysis, a general approach for treating ambiguities in so–called *ambiguous data sets*. After sketching an outline of variant analysis, we apply it together with a coin–tossing model to the problem of multiple local string alignment in protein sequences. The resulting new algorithm is shown to detect the target proteins, the recognition motif, and the association sites, also in a contaminated environment.

1 Introduction

Modern biology and biochemistry is, among other things, concerned with analyzing protein function and their interaction with polynucleotides and with other proteins. Besides the biological experiment such as phage display, SPOT analysis, NMR spectroscopy, fluorescence titration, statistical and computational methods based on the huge amount of sequence information available have been shown in the last fifteen years to be effective tools for this purpose. To this end, advanced statistical and computational methods are employed. We mention in this respect algorithms such as the EM algorithm and Gibbs sampling, and statistical models such as the HMM. They allow to determine transcription factor binding sites and their motifs, see Lawrence and Reilly (1990), Lawrence et al. (1993), Liu (1994), Liu et al. (1995). Gallegos and Ritter (2006) showed that variant analysis, a general statistical theory for treating ambiguities, could be applied for this purpose, too. As an example we used the binding sites for the cyclic AMP receptor protein (CRP) in the promoters of 18 genes of Escherichia coli, Stormo and Hartzell (1989), a protein–DNA interaction. We show in this note that the method is also effective in the analysis of protein–protein interactions. As an example, we study the recognition sites of two GYF domain–containing proteins of yeast, SMY2 and SYH1, see Kofler et al. (2005).

2 Parameter estimation under ambiguity and contamination

Contrary to statistics, pattern recognition deals with objects such as stars, plants, images, or signals rather than with data vectors as input to an analysis. In general, feature vectors are extracted from these objects and subsequently analyzed by statistical or other methods. It often happens that the extraction depends on an interpretation of the object which may not be unique in this stage of analysis. In this case we face ambiguity, a phenomenon that occurs in particular if the features are extracted by a machine. In the case of several reasonable, but not necessarily correct, interpretations, one may extract several feature sets of each object, one for each reasonable

obser_1	5.37	1.62	2.45	object_1	2.37	3.62	4.41
obser_2	4.11	2.21	2.13	object_1	1.14	1.21	3.12
obser_3	3.34	4.54	5.46	object_2	3.30	5.62	7.33
obser_4	8.35	6.76	7.78	object_2	8.11	6.29	7.13
obser_5	1.36	2.48	1.41	object_2	3.11	4.21	3.13
obser_6	5.76	7.61	2.15	object_3	6.54	5.22	8.46
	(a)				(b)		

Table 1: Three–dimensional data sets, (a) classical and (b) ambiguous. Whereas the six observations in data set (a) belong to six different objects, the six observations in data set (b) are extracted from three objects. Only one of the two variants of object 1 and one of the three variants of object 2 is a valid representative of its object. All variants of the same object are at first equally important.

interpretation. This means that each object may be represented by several distinct rows in the data set, called *variants* of the object, Ritter (2000). The variant that corresponds to the correct interpretation is the *regular* variant, the others are *irregular*. We call such a data set *ambiguous*, see Table 1. In some cases, some or all objects may possess more than one regular variant.

All questions about classical data sets may also be asked about ambiguous data sets – parameter estimation, discriminant analysis, clustering, ... An obstacle in the analysis of an ambiguous data set is the fact that it should be based on the regular variant of each object so that while carrying out the primary task the regular variant has to be estimated at the same time. This program was carried out for discriminant analysis, Ritter and Gallegos (2000), pure selection of the (or a) regular variant, Ritter and Gallegos (2002), and parameter estimation, Gallegos and Ritter (2006). Applications to image processing appear in Ritter and Pesch (2001), Ritter and Schreib (2000, 2001), Ritter and Gallegos (2000), and an application to motif discovery in genetics in Gallegos and Ritter (2006).

2.1 A statistical model of variants

Let $x_i = (x_{i,1}, \cdots, x_{i,h_i}, \cdots, x_{i,b_i})$, $x_{i,k}$ in some sample space E, represent the b_i variants of object $i \in 1..n$. For the purpose of this note, E is finite. Let us distinguish between two cases:

(i) At least one of the b_i variants of object i is regular, say h_i. We then assume that x_{i,h_i} is distributed according to some distribution with density f_γ for γ in some parameter space and call i a *regular object*.

(ii) None of the b_i variants is regular. In this case, object i is an *outlier*.

We assume that there are $r \leq n$ regular objects. The choice of the "correct" number is made later by validating the results obtained for various values of r. Central to variant analysis is the notion of a (variant) *selection* $\mathbf{h} = (h_1, \cdots, h_n)$, $h_i \in 0..b_i$. The relation $h_i = 0$ specifies object i as an outlier, whereas $h_i > 0$ means that h_i is the site of its regular variant. Thus, a selection implicitly contains the information about the regular objects in the data set. For example, if $r = 2$, ([object_1, 1], [object_2, 3], [object_3, 0]) is a selection in the data set (b) of Table 1. This selection considers objects 1 and 2 as regular and object 3 as an outlier. Our main objective is estimating the "true" selection and, thereby, the parameter γ.

Denote the cross section $(x_{i,h_i})_{h_i>0}$ specified by a selection $\mathbf{h}$ by $x_{\mathbf{h}}$; it is a classical data set with one row per regular object i. The cross section in Table 1 of the selection above is

object_1	2.37	3.62	4.41
object_2	3.11	4.21	3.13

Given i.i.d. random variables X_i, $1 \leq i \leq r$, distributed according to some unknown "true" distribution μ, the arithmetic means

$$\frac{1}{r}\sum_{i=1}^{r} -\ln f_\mu(X_i) \quad \text{and} \quad \frac{1}{r}\sum_{i=1}^{r} \ln \frac{f_\mu}{f_\gamma}(X_i)$$

converge to the entropy $-\mathrm{E}\ln f_\mu(X_1)$ of μ and to the Kullback–Leibler divergence $\mathrm{E}\ln \frac{f_\mu}{f_\gamma}(X_1)$ of μ and γ, respectively, P–a.s.. Hence, given a finite sequence $x_1, \ldots, x_r$ of observations, the means

$$\frac{1}{r}\sum_{i=1}^{r} -\ln f_\mu(x_i) \quad \text{and} \quad \frac{1}{r}\sum_{i=1}^{r} \ln \frac{f_\mu}{f_\gamma}(x_i)$$

are sample versions of these quantities. Neither of the two can be computed since f_μ is unknown, but their sum $\frac{1}{r}\sum_{i=1}^{r} -\ln f_\gamma(x_i)$ is an expression of γ alone. Two desirable aims are small entropy and small Kullback–Leibler divergence. These aims can be simultaneously achieved by minimizing this sum over γ. In the context of irregular variants and outliers we minimize this sum also w.r.t. all variant selections arriving at the criterion

$$\operatorname*{argmin}_{\mathbf{h}} \min_{\gamma} \sum_{h_i>0} -\ln f_\gamma(x_{i,h_i}). \tag{1}$$

Here, $\sum_{h_i>0} -\ln f_\gamma(x_{i,h_i})$ is the negative log-likelihood of γ for the regular variants w.r.t. the variant selection $\mathbf{h}$. The operation $\min_\gamma$ determines the m.l.e. $\gamma(\mathbf{h})$ of γ w.r.t these observations so that (1) may be rewritten as

$$\operatorname*{argmin}_{\mathbf{h}} \sum_{h_i>0} -\ln f_{\gamma(\mathbf{h})}(x_{i,h_i}).$$

Optimality of the method requires independence of all objects. Gallegos and Ritter (2006), Theorem 2.2, showed that Criterion (1) is the m.l.e. of γ and $\mathbf{h}$ w.r.t. a certain statistical model of the irregular variants that was called the *spurious–outliers model*.

Criterion (1) reduces the problem of estimating parameter and variant selection to minimizing the negative log-likelihood function of a distributional model and to a combinatorial optimization problem. Now, there are astronomically many selections, $\sum_{C\in\binom{1..n}{r}} \prod_{i\in C} b_i$; enumerating all is not feasible except for small instances and approximation algorithms are desirable. Such an algorithm is substantiated in the last–mentioned paper. Given a variant selection $\mathbf{h}$, define the negative estimated log-density

$$u_{\mathbf{h}}(i,k) = -\log f_{\gamma(\mathbf{h})}(x_{i,k}), \qquad k \in 1..b_i.$$

The basis of the algorithm is the following multi–point reduction step, a procedure that alternates parameter estimation and selection of the regular variants.

Multi–point reduction step

// Input: A selection $\mathbf{h}$;
// Output: A selection $\mathbf{h}_{\text{new}}$ with improved Criterion (1)
or the response "stop."

(i) Compute the estimate $\gamma(\mathbf{h})$;

(ii) for each object i, determine an element $h_{\mathrm{new},i} \in \mathrm{argmin}_{k \in 1..b_i} u_{\mathbf{h}}(i,k)$;

(iii) determine the r objects i with minimum values $u_{\mathbf{h}}(i, h_{\mathrm{new},i})$ and call the corresponding selection $\mathbf{h}_{\mathrm{new}}$;

(iv) *if* $u_{\mathbf{h}}(i, h_{\mathrm{new},i}) < u_{\mathbf{h}}(i, h_i)$ for at least on i then return $\mathbf{h}_{\mathrm{new}}$;
else "stop."

The multi–point reduction step is iterated until convergence. The variant selection obtained is self–consistent in the sense that it generates its original parameters. The optimal solution shares this property but the result of the iteration is not necessarily optimal. Therefore, the multistart method has to be applied to reduce the criterion at least to a low value.

Alternative methods for minimizing Criterion (1) are local search, the Metropolis–Hastings algorithm, the EM algorithm, and Gibbs sampling. However, to our experience the multipoint reduction step is competitive with these methods.

2.2 A coin–tossing model

An interesting and important special case is a discrete model with sample space $E = (1..\,s)^d$ where the regular variants are generated by tossing d independent, possibly biased, s–sided coins. The parameter γ is an $s \times d$ table $\mathbf{p}$ of real numbers $p_{y,m} \geq 0$ whose columns sum to 1, the *position–specific score matrix* PSSM. Each variant $x \in E$ generates a path in this table that visits each column exactly once. Its probability is the product of the entries along the path,

$$f_{\mathbf{p}}(x) = \prod_{m=1}^{d} p_{x_m,m}.$$

Let $\mathbf{h}$ be a selection and let $n_{y,m}(\mathbf{h}) = \#\{i \mid x_{i,h_i,m} = y\}$ be the frequency of the outcome y at position m taken over the r selected variants of length d. These frequencies sum up to rd. The m.l.e. of the PSSM consists of the relative frequencies $n_{y,m}(\mathbf{h})/r$, $y \in 1..s,\ m \in 1..d$, and, up to a multiplicative constant, the maximum value of the likelihood function

$$f_{\mathbf{p}}(x_{\mathbf{h}}) = \prod_{i:h_i \geq 1} \prod_{m=1}^{d} p_{x_{i,h_i,m},m} = \prod_{m=1}^{d} \prod_{y \in 1..s} p_{y,m}^{n_{y,m}(\mathbf{h})},$$

cf. (1), equals their negative entropy.

The likelihood may be optimized by multistart replication of the iterative application of multipoint reduction steps. In the present context, the quantities $u_{\mathbf{h}}(i,k)$ become

$$u_{\mathbf{h}}(i,k) = -\sum_{m} \ln \frac{n_{x_{i,k,m},m}(\mathbf{h})}{r}.$$

When, in item (ii) of the multi–point reduction step, a new "regular" variant is selected for object i, the relative frequencies $n_{y,m}/r$ are biased towards its current regular variant. Therefore, replacing the relative frequencies appearing in $u_{\mathbf{h}}$ with $n'_{y,m}/(r-1)$ offers a big advantage, the prime indicating omission of this object. Note that the probability estimates are now based on $r-1$ observations. Therefore, instead of the maximum likelihood, Laplace's Law of Succession should be used for estimating the probabilities $p_{y,m}$ which means that the numbers $(n'_{y,m}+1)/(r-1+s)$ replace the relative frequencies $n'_{y,m}/(r-1)$. In the extreme case of a data set consisting of one line one has the unbiased prior $1/s$.

3 Study of a protein–protein interaction

GYF (glycine–tyrosine–phenylalanine) domains are highly conserved protein domains expressed in human (PERQ2), yeast (SMY2 and its paralog YPL105C), and plant (GYN4), see Kofler et al. (2005). They are characterized by two beta strands, an extended loop in between, and a successive alpha helix which is flanked by the patterns GPF (glycine–proline–phenylalanine) and GYF. A GYF domaine is known to recognize proline–rich patterns in targets, the common recognition signature being PPG (proline–proline–glycine), Kofler et al. (2005).

Is it possible to detect this signature by variant analysis? More precisely: does the algorithm detect short segments in the target polypeptides which approximately match each other pairwise? The answer to this problem of *multiple local string alignment*, see e.g. Gusfield (1997), is yes. The signature is detected in a set of protein sequences that may even be contaminated in the sense that an unknown subset of the proteins, only, act as targets. For this purpose, two data sets were compiled: Data Set A is heavily contaminated containing 100 targets and 100 non–targets, whereas data set B is moderately contaminated and contains the same 100 targets and the first 40 non–targets of Data Set A. The targets were taken from the supplemental material of Kofler et al. (2005), whereas the non–targets are the first 100 entries of the Stanford Saccharomyces Genome Database. The latter act as outliers in the data sets. We show that, despite the contamination, the interacting proteins, the common motif, and the sites where the interaction takes place can be discovered. Moreover, exact knowledge of the motif length is not necessary.

The data set consists of n polypeptides. Each polypeptide (= object) of length l amino acids gives rise to $l-d+1$ (overlapping) segments of length d, one for each possible initial site in the sequence. These are the variants of the object and E is the d–fold Cartesian product of the set of the 20 naturally occurring amino acids. An array of r initial sites is called an *alignment*. It corresponds to a variant selection. The remaining $n-r$ sequences are outliers, i.e. considered non-targets w.r.t. the alignment. Assuming the different polypeptides to be independent as in Sect. 2.1, one may apply the foregoing theory and the modified multi–point reduction step.

Almost all pairs of residues appear very often in most medium–size and long proteins for combinatorial reasons; they are insignificant. Therefore, the smallest length d considered is three residues. The runs with $d = 3, \cdots, 6$, $r = 80, 90, \cdots, 140$, and 100,000 replications of reduction–step iterations took between one and four hours, each.

The algorithm finds two motifs: accumulations of serines and the motif PPG, see Table 2. The former are ubiquitous and prevail in the heavily contaminated data set at the motif lengths five and six and in the moderately contaminated data set at length six. The latter prevails in the moderately contaminated data set up to length five and is known to be the consensus pattern of the recognition site of the GYF domain. Since the PPG motif was identified, the same was of course true for the association sites and the target proteins of the GYF domain. Table 2 confirms the well–known fact that it is harmful to assume too many regular objects. For the length four, the result of the heavily contaminated data set A breaks down under the assumption of more regular objects than there actually are (100). By contrast, the moderately contaminated data set resists the assumption of 130 regular objects even at length five.

The specificities of the motif PPG* are shown in Table 3. The highest specificities at the uncertain fourth position are assumed by the mostly hydrophobic side chains mentioned in the caption (an exception is alanine which is neutral).

The study shows that the method is to a certain extend robust against outliers and against an unfavorable choice of motif length and assumed number of outliers.

Acknowledgment. We thank Frau Saskia Nieckau for her implementation of the algorithm.

d / r	3	4	5	6
80	S S S S L L P P G	P P G *	S S S S S	S * * * S S
90	S S S P P G	P P G *	S * S S S	S * S * S S
100	P P G	P P G *	S * S S S	S * * * S S
110	P P G	S * S S	S * S S *	S * * * S *
120	P P G	S * S S	S * * S *	S * S * S *
130	P P G	S * S S	S * * S S	S * * * S *
140	S S S	S * S S	S * * S *	S * * * * *

80	P P G	P P G *	* P P G *	S * S * * *
90	P P G	P P G *	* P P G *	S * S * * *
100	P P G	P P G *	* P P G *	S * * * * *
110	P P G	P P G *	* P P G *	S * * * * *
120	P P G	P P G *	* P P G *	S * * * * *
130	P P G	P P G *	* P P G *	S * * * * *
140	P P G	P P G *	S * * S *	S * * * * *

Table 2: Multiple sequence alignment with various motif lengths d and assumed numbers r of regular elements. Top: motifs found in a data set of 100 positive and 100 negative sequences (Data Set A), bottom: 100 positive and 40 negative sequences. Residues with specificities $\geq 60\%$ at a position are shown. The lack of such a highly significant residue at some site is indicated by an asterisk.

res / pos	A	C	D	E	F	G	I	L	M	P	R	S	V	W	Y
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.06	0.04	0.01	0.01	0.09	0.03	0.21	0.15	0.02	0.01	0.01	0.01	0.17	0.05	0.13

Table 3: Transposed of the PSSM for the Data Set B with parameters $r = 100$ and $d = 4$. Only residues with positive specificities at some position are shown. The consensus sequence is PPG*, the uncertain last position being mainly occupied by the hydrophobic residues isoleucine, valine, leucine, tyrosine, phenylalanine, and alanine.

References

Gallegos, M. T., and Ritter, G. Parameter estimation under ambiguity and contamination with the spurious model. *J. Multivariate Analysis 97* (2006), 1221–1250.

Gusfield, D. *Algorithms on Strings, Trees, and Sequences; Computer Science and Computational Biology.* Cambridge University Press, Cambridge, 1997.

Kofler, M., Motzny, K., and Freund, C. GYF domain proteomics reveals interaction sites in known and novel target proteins. *Molecular & Cellular Proteomics 4* (2005), 1797–1811. http://www.mcponline.org.

Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., and Wootton, J. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science 262* (1993), 208–214.

Lawrence, C. E., and Reilly, A. A. An expectation maximizatiom (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *PROTEINS: Structure, Function, and Genetics 7* (1990), 41–51.

Liu, J. S. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *JASA 89* (1994), 958–966.

Liu, J. S., Neuwald, A. F., and Lawrence, C. E. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *JASA 90* (1995), 1156–1169.

Ritter, G. Classification and clustering of objects with variants. In *Data Analysis, Scientific Modeling and Practical Application*, W. Gaul, O. Opitz, and M. Schader, Eds., Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg, 2000, pp. 41–50.

Ritter, G., and Gallegos, M. T. A Bayesian approach to object identification in pattern recognition. In *Proceedings of the 15th International Conference on Pattern Recognition* (Barcelona, 2000), A. S. et al., Ed., vol. 2, pp. 418–421.

Ritter, G., and Gallegos, M. T. Bayesian object identification: variants. *Journal of Multivariate Analysis 81* (2002), 301–334.

Ritter, G., and Pesch, C. Polarity–free automatic classification of chromosomes. *Computational Statistics and Data Analysis 35* (2001), 351–372.

Ritter, G., and Schreib, G. Profile and feature extraction from chromosomes. In *Proceedings of the 15th International Conference on Pattern Recognition* (Barcelona, 2000), A. S. et al., Ed., vol. 2, pp. 287–290.

Ritter, G., and Schreib, G. Using dominant points and variants for profile extraction from chromosomes. *Patt. Rec. 34* (2001), 923–938.

Stormo, G. D., and Hartzell III, G. W. Identifying protein–binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. 86* (1989), 1183–1187.