ELSEVIER

Mechanisms of Ageing and Development 124 (2003) 109–114

mechanisms of ageing and development

www.elsevier.com/locate/mechagedev

Integrated analysis of transcript profiling and protein sequence data

L.R. Grate [a], C. Bhattacharyya [b], M.I. Jordan [b,c], I.S. Mian [a,*]

[a] *Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[b] *Division of Computer Science, University of California Berkeley, Berkeley, CA 94720, USA*
[c] *Department of Statistics, University of California Berkeley, Berkeley, CA 94720, USA*

Abstract

Transcript profiling can be used to elucidate the molecular and cellular mechanisms involved in ageing and cancer. A recent study of human gastrointestinal stromal tumours (GISTs) with mutations in the *KIT* gene, Cancer Res. 61 (2001) 8624 exemplifies a common type of investigation. cDNA microarrays were used to generate measurements for 1987 clones in two types of tissues: 13 *KIT* mutation-positive GISTs and 6 spindle cell tumours from locations outside the gastrointestinal tract. Statistical problems associated with such two-class, high-dimensional profiling data include simultaneous classification and relevant feature identification, probabilistic clustering and protein sequence family modelling. Here, the GIST data were reexamined using specific solutions to these problems, namely sparse hyperplanes, naïve Bayes models and profile hidden Markov models respectively. The integrated analysis of molecular profiling and sequence data highlighted 6 clones that may be of clinical and experimental interest. The protein encoded by one of these putative biomarkers defined a novel protein family present in diverse eucarya. The family may be involved in chromosome segregation and/or stability. One family member is a potential biomarker identified recently from a retrospective analysis of transcript profiles for sporadic breast cancer samples from patients with poor and good prognosis, Signal Process. (in press).

Keywords: Sparse hyperplanes (L_1 norm minimisation); Probabilistic clustering; Hidden Markov models; Cancer; Chromosome segregation

1. Introduction

Molecular profiling technologies provide a snapshot of the approximate inventory of transcripts, proteins, metabolites or other species present in biological samples of interest. Such studies herald an opportunity to examine molecular function(s), tasks performed by individual gene products, in the context of (sub)cellular processes performed by molecular ensembles. When extended to the tissue, organ and organism levels and incorporated with the spatial and temporal locations of molecules, such endeavours will be keystones of efforts to elucidate how and why the physiological vigour of an organism declines over time.

Monitoring transcript levels in tissues from young and old mice, for example, could assist in enunciating the genes and possibly processes associated with age-related changes. Such a study would be analagous to a recent investigation of gastrointestinal tissues that generated transcript profiles for 1987 clones in 19 samples categorised as *KIT* mutation-positive gastrointestinal stromal tumours (GISTs) or spindle cell tumours (Allander et al., 2001). Here, this published exemplar will be used to demonstrate the utility of specific statistical techniques for solving some of the analytical challenges posed by two-class, high-dimensional molecular profiling data. More generally, this retrospective computational study reveals the ability of a corpus in the public domain to yield unanticipated observations and to generate new predictions of widespread interest.

Given two-class data, analytical tasks include classification and prediction, relevant feature identification and clustering. Classification and prediction methods estimate a model from data belonging to known categories and use the learned system to assign the (unknown) class of a new data point. Thus, a classifier trained using the 13 GIST and 6 spindle cell tumour samples could be used to create a clinical decision support system designed to predict the tumour type of a new patient sample. Relevant feature identification

* Corresponding author.
E-mail address: smian@lbl.gov (I.S. Mian).

0047-6374/03/$ - see front matter
doi:10.1016/S0047-6374(02)00174-4

methods define which features in the data best differentiate classes. Enumerating a subset of the 1987 clones which distinguished the two tumour types could delineate robust and reliable targets for intervention, diagnosis and imaging ('biomarkers'). Clustering methods find groups of data points with similar patterns. Discovering which of the 1987 clones had similar expression patterns across the 19 samples could suggest genes with potentially common properties and/or non-coding (regulatory) regions.

As shown recently (Moler et al., 2000a), comparative molecular profile analysis and comparative sequence analysis provide complementary biological insights. Sequence analysis methods characterise the domain structure and/or function of proteins. For clones such as those distinguishing GISTs from spindle cell tumours, comprehensive annotation of the encoded proteins could enhance knowledge of molecular mechanisms and genetic networks important in these tumours.

For the GIST two-class, high-dimensional transcript profiling data (Allander et al., 2001), the analytical problems outlined above were solved using extant statistical techniques and specific software implementations. These were (i) simultaneous classification and relevant feature identification via estimation of a sparse hyperplane (the programme LIKNON (Bhattacharyya et al., in press)), (ii) probabilistic clustering via estimation of a naïve Bayes model (AUTOCLASS (Cheeseman et al., 1995)), and (iii) protein sequence family modelling via estimation of a profile hidden Markov model (HMM) (SAM (Hughey and Krogh, 1995)). Previously, these tools were utilised to analyse a variety of transcript and protein profiles (Bhattacharyya et al., in press, Moler et al., 2000a,b; Chow et al., 2001) and proteins associated with ageing (Mian, 1998; Huang et al., 1998; Lim et al., 2001). Here, they were applied as an ensemble to a single data set. This qualitative fusion of heterogeneous techniques generated novel hypotheses about molecular and physiological gene function and disease aetiology. In particular, one of the 6 distinguishing clones defined a new, phylogenetically conserved protein family with a possible role in chromosome segregation and/or stability.

2. Methods

2.1. Two-class transcript profiling data

Transcript profiles from cDNA microarray experiments performed using 19 samples (Allander et al., 2001) were downloaded from: http://www.nhgri.nih.gov/dir/microarray/gist_data.txt and used as is. For each sample, the 1987-dimensional vectors consisted of calibrated ratios for clones in the sample of interest compared to a common reference cell line, OsA–Cl. There were two types of samples, 13 labelled as *KIT* mutation-positive GISTs and 6 as spindle cell tumours.

The transcript profiles can be viewed as a 1987×19 matrix. Rows will be termed gene profile vectors and columns sample profile vectors. Columns are labelled on the basis of tumour type (GIST or spindle cell tumour).

2.2. Simultaneous classification and relevant feature identification

The programme LIKNON (Bhattacharyya et al., in press) implements a particular strategy for simultaneous classification and relevant feature identification. It constructs a sparse hyperplane by minimising an L_1 norm and uses linear programming to solve the underlying optimisation problem. The input was 19 1987-dimensional labelled sample profile vectors. In one pass through this two-class data set, LIKNON defined a classifier and a small number of relevant features (clones able to distinguish GISTs from spindle cell tumours).

The ability of the LIKNON classifier to discriminate between GISTs and spindle cell tumours using only the six relevant features was assessed by computing the leave-one-out error. This error is a proxy for generalisation performance when there are few examples. The 19 6-dimensional example vectors were divided into an estimation set of 18 samples and a test set formed by the withheld sample. LIKNON was used to predict the class of the test sample. This estimation and evaluation procedure was repeated 19 times so that the class of each sample was assigned by a classifier estimated using all other 18 samples. The leave-one-out error is the number of test samples whose predicted class differs from the known class (maximum 19).

2.3. Probabilistic clustering

The programme AUTOCLASS (Cheeseman et al., 1995) implements a particular strategy for probabilistic clustering. It estimates a naïve Bayes model and uses a Bayesian approach to discover (automatically) the numbers of clusters K which best fit the data. The input was 1987 19-dimensional unlabelled gene profile vectors. The resultant model was characterised by $K \times 19$ distinct probability distributions i.e. for gene cluster k, there was a Gaussian specifying transcript levels in each of the 19 samples. Although clones can have partial class membership, the trained model was used to assign each of the 1987 gene profile vectors to a single cluster (hard assignment).

2.4. Protein sequence analysis

The programme SAM (Hughey and Krogh, 1996) implements a particular strategy for sequence family modelling. It estimates an HMM for a set of sequences

believed to be similar and can incoporate prior knowledge into the model building process. The protein encoded by one of the six relevant features was annotated as an open reading frame (ORF). Sequence homologues of this ORF were identified by using it as the query for the programme PSI-BLAST (Altschul et al., 1997). The input to SAM was the ensuing, newly defined set of related proteins. The resultant HMM was employed to generate a multiple sequence alignment of this protein family.

2.5. Qualitative fusion of heterogeneous techniques

When using sample profile vectors to define relevant features, LIKNON does not take into consideration the groupings of gene profile vectors. When employing gene profile vectors to cluster clones, AUTOCLASS does not take into consideration the existence of two types of samples. HMMs only model a user-defined set of sequences. The results from application of these independent approaches were integrated as follows. Gene profile vector clusters containing LIKNON relevant features were identified. Other clones in these clusters were examined. Predictions about the function of a distinguishing clone designated as a hypothetical protein were formulated using database searching and a custom built HMM.

3. Results

The six distinguishing clones defined here represent a tractable number of biomarkers for possible eventual clinical deployment. Simultaneous classification and relevant feature identification was performed using LIKNON (Bhattacharyya et al., in press). Six of the 1987 clones assayed in 13 GISTs and 6 spindle cell tumours were defined as being able to discriminate between the two types of samples. The leave-one-out error of these six relevant features was zero out of 19. This good discriminatory power suggests a potential to generalise well i.e. to predict accurately whether a new patient sample was a GIST or a spindle cell tumour. These six clones should be regarded as a small, though not necessarily unique set of relevant features because other clone subsets could distinguish GISTs from spindle cell tumours equally well (Bhattacharyya et al., in press and Chow et al., 2001).

High-density tissue microarray (TMA) technology use immunohistochemical analysis of protein expression in large numbers of clinical specimens to facilitate rapid translation of findings from molecular profiling studies to clinical specimens (Kononen et al., 1998). The in vivo importance of the proteins encoded by the 6 distinguishing clones requires validation by TMAs. The clones were annotated as G protein-coupled receptor 20 (Clone ID 2568905), protein kinase C θ (2164126), hypothetical protein FLJ10261 (742679), ESTs (1686218), complement component 1, q subcomponent, β polypeptide (85128) and matrix Gla protein (590264).

Clones in the same gene profile vector clusters as LIKNON relevant features proffer candidates for interrogating molecular mechanisms and networks. Probabilistic clustering was performed using a naïve Bayes model as implemented by AUTOCLASS (Cheeseman et al., 1995). The model grouped the 1987 gene profile vectors into 26 classes. One class, designated Cluster 19, contained all six distinguishing clones as well as the *KIT* gene itself. For Cluster 19, the mean (standard deviation) of the Gaussian modelling the calibrated ratio of clones in the 13 GIST samples was in the range 2.39–3.98 (1.38–0.86). For spindle cell tumours, the corresponding values were 0.958–1.93 (1.22–1.55). Thus, this class exhibits a pattern of expression that differs in the two tumour types. Additional investigation of the 52 clones assigned to Cluster 19 should provide insights into how and why GISTs differ from spindle cell tumours. Table 1 lists these clones. They include genes involved in signal transduction, neuropeptides, the complement system and so on.

Characterisation of the protein encoded by a distinguishing clone can suggest subsequent targeted experiments. Cluster 19 contained a number of unannotated clones including a distinguishing clone annotated as 'hypothetical protein FLJ10261'. Protein sequence analysis and family modelling using PSI-BLAST and HMMs indicated that FLJ10261 defined a previously unknown and uncharacterised phylogenetically conserved and diverse protein family. Fig. 1 shows an HMM-generated multiple sequence of the family. Of the 16 eucaryotic sequences shown, 15 are unannotated in terms of a putative molecular function. The exception, DmAxs, is a fly protein annotated as being involved in aberrant X segregation. Preliminary results suggest that members of this conserved family may be transmembrane proteins.

Family members from model organisms should prove useful for investigating the molecular, cellular, tissue, organ and organismal phenotypes of normal and aberrant forms of the human homologues. Conserved positions in the multiple sequence alignment represent good candidates for site-directed mutatgenesis studies. Cellular imaging of such mutants would enable any consequences on chromosome and nuclear localisation to be ascertained. Comparison with a wild type organism using transcript profiling technology could suggest downstream targets.

Given the known assocation between chromosome segregation and cancer (Jallepalli and Lengauer, 2001), members of this protein family may be of widespread biological and clinical interest. In addition to human FLJ10261, a second human protein shown in Fig. 1 has

Table 1
The 52 clones assigned to Cluster 19, one of 26 clusters estimated from the 1987 19-dimensional gene profile vectors

Clone id	Description and class probability	
269806	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	1.000
2568905	G protein-coupled receptor 20‡	1.000
2568905	G protein-coupled receptor 20	1.000
205239	Protein kinase C, theta	1.000
2164126	Protein kinase C, theta‡	1.000
590264	Matrix Gla protein‡	1.000
375827	Protein tyrosine phosphatase type IVA, member 3	1.000
434833	Erythrocyte membrane protein band 4.1 (elliptocytosis 1, RH-linked)	0.995
303139	Homo sapiens cDNA FLJ14054 fis, clone HEMBB1000240	1.000
970731	Phosphodiesterase 4C, cAMP-specific (dunce (Drosophila)-homolog phosphodiesterase E1)	0.967
67769	Potassium channel, subfamily K, member 3 (TASK)	0.846
2551468	Guanylate cyclase 1, soluble, alpha 3	1.000
160485	Guanylate cyclase 1, soluble, alpha 3	1.000
2569769	Carbonic anhydrase II	1.000
2017204	Proenkephalin	1.000
878836	Secretory granule, neuroendocrine protein 1 (7B2 protein)	0.868
2469213	Annexin A3	1.000
2514318	MyoD family inhibitor	0.999
461351	Transcription factor 21	1.000
1635596	Ras association (RalGDS/AF-6) domain family 2	1.000
898305	Neuroblastoma, suppression of tumorigenicity 1	0.999
344720	Glycophorin C (Gerbich blood group)	0.927
153505	Dermatopontin	1.000
207274	Insulin-like growth factor 2 (somatomedin A)	0.993
296448	Insulin-like growth factor 2 (somatomedin A)	1.000
207274	Insulin-like growth factor 2 (somatomedin A)	1.000
366541	Chymotrypsin-like	1.000
2562939	Serine (or cysteine) proteinase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 1	1.000
2782586	Hemoglobin, alpha 2	0.999
2461206	Hemoglobin, beta	0.995
203732	Fibrinogen-like 2	1.000
878182	Alpha-2-macroglobulin	1.000
770858	CD34 antigen	0.988
868652	Complement component 4A	1.000
2559389	Complement component 4A	0.672
1957039	Complement component 4B	1.000
898122	Complement component 7	1.000
2569884	Complement component 1, s subcomponent	1.000
85634	Complement component 1, s subcomponent	1.000
85128	Complement component 1, q subcomponent, beta polypeptide‡	1.000
2419934	Down syndrome critical region gene 1-like 1	1.000
1686218	ESTs‡	1.000
742679	Hypothetical protein FLJ10261‡	1.000
1161564	KIAA0353 protein	0.977
813603	KIAA1075 protein	1.000
70218	Homo sapiens mRNA; cDNA DKFZp434O1616 (from clone DKFZp434O1616)	0.865
212542	Homo sapiens cDNA FLJ12900 fis, clone NT2RP2004321	1.000
491519	Homo sapiens clone 24775 mRNA sequence	0.999
486493	DKFZP434C211 protein	1.000
767181	DKFZP564G202 protein	1.000
767181	DKFZP564G202 protein	1.000
136070	DKFZP586A0522 protein	1.000

Clones marked with ‡ are LIKNON relevant features for 13 GISTs and 6 spindle cell tumours. Class probability indicates the probability of the gene profile vector for Cluster 19. When this value is less than 1.0, the gene has partial membership in one or more of the other 25 classes.

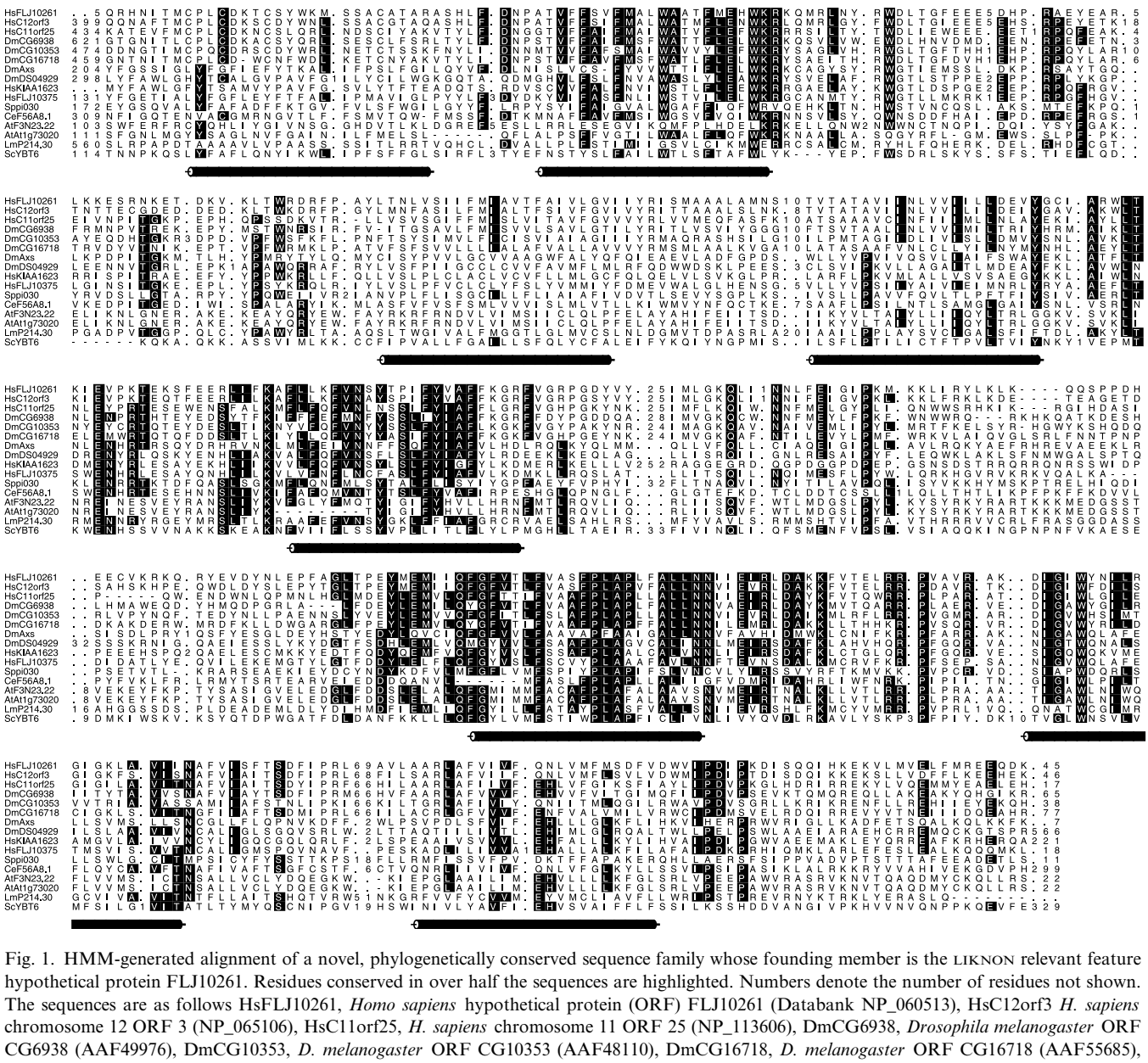

Fig. 1. HMM-generated alignment of a novel, phylogenetically conserved sequence family whose founding member is the LIKNON relevant feature hypothetical protein FLJ10261. Residues conserved in over half the sequences are highlighted. Numbers denote the number of residues not shown. The sequences are as follows HsFLJ10261, *Homo sapiens* hypothetical protein (ORF) FLJ10261 (Databank NP_060513), HsC12orf3 *H. sapiens* chromosome 12 ORF 3 (NP_065106), HsC11orf25, *H. sapiens* chromosome 11 ORF 25 (NP_113606), DmCG6938, *Drosophila melanogaster* ORF CG6938 (AAF49976), DmCG10353, *D. melanogaster* ORF CG10353 (AAF48110), DmCG16718, *D. melanogaster* ORF CG16718 (AAF55685), DmAxs, *D. melanogaster* aberrant X segregation protein Axs-P2 (NP_476777), DmDS04929, *D. melanogaster* ORF DS04929.1 (AE003406_125), HsKIAA1623, *H. sapiens* ORF KIAA1623 (XP_050644), HsFLJ10375, *H. sapiens* ORF FLJ10375 (NP_060545), Sppi030, *Schizosaccharomyces pombe* ORF pi030 (BAA21409), CeF56A8.1, *Caenorhabditis elegans* ORF F56A8.1 (T22762), AtAt1g73020, *Arabidopsis thaliana* ORF At1g73020 (NP_177445), AtF3N23.22, *A. thaliana* ORF F3N23.22 (AC008017_22), LmP214.30, *Leishmania major* ORF P214.30 (CAC14637), ScYBT6, *Saccharomyces cerevisiae* ORF (YBT6_YEAST). Cylinders represent potential transmembrane helices for ScYBT6 and are those defined in the database entry.

a role in cancer. ORF3 on human chromosome 12 (HsC12orf3) is a LIKNON relevant feature for a two-class sporadic breast carcinoma transcript profiling data set consisting of 46 patients with distant metastases < 5 years and 51 patients with no distant metastases $\geqslant 5$ years (Bhattacharyya et al., in press).

4. Discussion

The qualitative fusion of heterogeneous information and statistical techniques outlined here provides a starting point for integrating knowledge across a range of spatial and temporal resolutions. Data driven model-

ling of a published corpus resulted in the formulation of specific predictions relevant to the biology of two types of tumours. These hypotheses can stimulate the design of new experimental studies aimed at validating and extending the findings. The identification of a new protein family which contains members that differentiate not only GISTs from spindle cell tumours, but also sporadic breast carcinomas from patients with good and poor prognoses highlight the importance of DNA, chromatin, chromosome and nuclear stability, organisation and architecture. Whether any of the proteins shown in Fig. 1 have roles in the ageing process remains to be seen.

Although profiling and sequence technologies generate data that are complementary in terms of the biological knowledge they can provide, they will be insufficient in and of themselves to ascertain the complex processes underlying how and why organisms age. The value of the strategy discussed here is that it specifies a foundation for probing the means by which the functions of putative biomarkers are coordinated individually and as a collective, and how interactions in space and time contribute to the normal and disease state. An outstanding challenge is developing a formal statistical framework and attendant software to augment and drive such experimental investigations.

Acknowledgements

This work was supported by NSF grant IIS-9988642, the Director, Office of Energy Research, Office of Health and Environmental Research, Division of the US Department of Energy under Contract No. DE-AC03-76F00098 and an LBNL/LDRD through US Department of Energy Contract No. DE-AC03-76SF00098.

References

Allander, S., Nupponen, N., Ringner, M., Hostetter, G., Maher, G., Goldberger, N., Chen, Y., Elkahloun, C.J.A., Meltzer, P., 2001. Gastrointestinal stromal tumors with KIT mutations exhibit a remarkably homogeneous gene expression profile. Cancer Research 61, 8624–8628.

Altschul, S., Madden, T., Schaffer, A., Zhang ,J., Zhang, Z., Miller, W., Lipman, D., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research 25, 3389–3402, the www-interface at the NCBI is available at: http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast.

Bhattacharyya, C., Grate, L., Rizki, A., Radisky, D., Molina, F., Jordan, M., Bissell, M., Mian, I., Simultaneous relevant feature identification and classification in high-dimensional spaces: application to molecular profiling data, Signal Processing, in press.

Cheeseman, P., Stutz, J., 1995. Bayesian Classification (AUTOCLASS): theory and Results, in: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data miNing, AAAI Press/MIT Press, pp. 153–180, the software is available at: http://www.gnu.org/directory/autoclass.html.

Chow, M., Moler, E., Mian, I., 2001. Identifying marker genes in transcription profile data using a mixture of feature relevance experts. Physiological Genomics 5, 99–111.

Huang, S., Li, B., Gray, M., Oshima, J., Mian, I., Campisi, J., 1998. The premature aging syndrome protein, Wrn, is a 3′ to 5′ exonuclease. Nature Genetics 20, 114–116.

Hughey, R., Krogh, A., 1995. SAM: sequence alignment and modelling system software, Technical Report UCSC-CRL-95-7, University of California, Santa Cruz, Computer and Information Sciences Department, Santa Cruz, CA 95064, the software is available at: http://www.soe.ucsc.edu/research/compbio/sam.html.

Hughey, R., Krogh, A., 1996. Hidden Markov models for sequence analysis: extension and analysis of the basic method, CABIOS 12, 95–107, The hidden Markov model software can be accessed at: http://www.cse.ucsc.edu/research/compbio/sam.html.

Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M., Sauter, G., Kallioniemi, O.P., 1998. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nature Medicine 4, 844–847.

Jallepalli, P., Lengauer, C., 2001. Chromosome segregation and cancer: cutting through the mystery. Nature Review Cancer 1, 109–117.

Lim, C.-S., Mian, I., Dernburg, A., Campisi, J., 2001. A new regulator of telomere metabolism in *C. elegans*, encoded by the life span regulator *clk-2*. Current Biology 11, 1706–1710.

Mian, I., 1998. Sequence, structural, functional and phylogenetic analysis of three glycosidase families. Blood Cells, Molecules and Disease 24, 83–100.

Moler, E., Chow, M., Mian, I., 2000a. Analysis of molecular profile data using generative and discriminative methods. Physiological Genomics 4, 109–126.

Moler, E., Radisky, D., Mian, I., 2000b. Integrating naïve Bayes models and external knowledge to examine copper and iron homeostasis in *Saccharomyces cerevisiae*. Physiological Genomics 4, 127–135.