A Network-QSAR model for prediction of genetic-component biomarkers in Human Colorectal Cancer

Santiago Vilar, Humberto González-Díaz, Lourdes Santana, Eugenio Uriarte

▶ **To cite this version:**

Santiago Vilar, Humberto González-Díaz, Lourdes Santana, Eugenio Uriarte. A Network-QSAR model for prediction of genetic-component biomarkers in Human Colorectal Cancer. Journal of Theoretical Biology, 2009, 261 (3), pp.449. 10.1016/j.jtbi.2009.07.031 . hal-00559149

HAL Id: hal-00559149
https://hal.science/hal-00559149

Submitted on 25 Jan 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Author's Accepted Manuscript

www.elsevier.com/locate/yjtbi

A Network-QSAR model for prediction of genetic-component biomarkers in Human Colorectal Cancer

Santiago Vilar, Humberto González-Díaz, Lourdes Santana, Eugenio Uriarte

PII: S0022-5193(09)00345-2
DOI: doi:10.1016/j.jtbi.2009.07.031
Reference: YJTBI 5650

To appear in: *Journal of Theoretical Biology*

Received date: 15 March 2009
Revised date: 20 July 2009
Accepted date: 25 July 2009

Cite this article as: Santiago Vilar, Humberto González-Díaz, Lourdes Santana and Eugenio Uriarte, A Network-QSAR model for prediction of genetic-component biomarkers in Human Colorectal Cancer, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2009.07.031

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

A Network-QSAR Model for Prediction of Genetic-Component Biomarkers in Human Colorectal Cancer

Santiago Vilar, Humberto González-Díaz, Lourdes Santana and Eugenio Uriarte*

Department of Organic Chemistry, Faculty of Pharmacy, and Unit of Bioinformatics & Connectivity Analysis of Systems (UBICAS), Institute of Industrial Pharmacy , University of Santiago de Compostela, Santiago de Compostela 15782, Spain

* Corresponding author. Fax: +34-981-594912; e-mail: eugenio.uriarte@usc.es

Accepted manuscript

Abstract

The combination of the network theory and the calculation of Topological Indices (TIs) allow establishing relationships between the molecular structure of large molecules like the genes and proteins and their properties at a biological level. This type of models can be considered Quantitative Structure-Activity Relationships (QSAR) for biopolymers. In the present work a QSAR model is reported for proteins, related to Human Colorectal Cancer (HCC) and codified by different genes that have been identified experimentally by Sjöblom *et al*.(Science, 314 (2006) 268-274) among more than 10000 human genes. The 69 proteins related to Human Colorectal Cancer (HCCp) and a control group of 200 proteins not related to HCC (no-HCCp) were represented through a HP Lattice type Network. Starting from the generated graphs we calculate a set of descriptors of electrostatic potential type (ξk) that allow to establish, through a Linear Discriminant Analysis (LDA), a QSAR model of relatively high percentage of good classification (higher than 80 %) to differentiate between HCCp and no-HCCp proteins. The purpose of this study is helping to predict the possible implication of a certain gene and/or protein (biomarker) in the colorectal cancer. Different procedures of validation of the obtained model have been carried out in order to corroborate its stability, including cross-validation series (CV) and evaluation of an additional series of 200 no-HCCp. This biostatistic

ACCEPTED MANUSCRIPT

methodology could be applied to predict Human Colorectal Cancer biomarkers and to understand much better the biological aspects of this disease.

Keywords: Protein Sequence; Colorectal Cancer; Markov Chains; HP Lattice; Complex Networks; Biomarkers; QSAR; Linear Discriminant Analysis; Sequence Alignment; Electrostatic Potential.

Accepted manuscript

ACCEPTED MANUSCRIPT

1. Introduction

The colorectal cancer consists of uncontrolled growths of abnormal cells in that part of the intestine (Boursi and Arber, 2007; Stein and Schlag, 2007). In the cancerous processes, the cells undergo transformations or mutations in the DNA and multiply very fast since they are not subject to the usual restrictions of cellular proliferation. These cells can invade and destroy the tissue around. If they enter the sanguineous or lymphatic system, they can spread to any part of the organism and produce damages in other organs. This process of spread is called metastasis.

Many colorectal cancers are thought to arise from adenomatous polyps in the colon. These cell growths are usually benign, but some may develop into cancer over time. More than 95% of the cancerous tumors of the colon and the rectum are adenocarcinomas. The majority of the time, the diagnosis of localized colon cancer is through colonoscopy. Therapy is usually through surgery, which in many cases is followed by chemotherapy. When the colorectal cancer is detected early, it can be frequently cured. The rate of mortality caused by this type of cancer has decreased over the last 20 years, probably because nowadays there are better treatments and many cases are detected in the early stage of the cancer. The colorectal cancer is very common in men and women all over the world. Its real causes are not known. The investigation has shown that people with certain risk factors, like age, diet, smoking or certain genetic alterations, have more probability than others to develop colorectal cancer (Luchtenborg et al., 2007; Schafmayer et al., 2007; Young, 2007). In this sense, the development of tools for search of Colorectal Cancer biomarkers becomes of the major importance (Yasui et al., 2003).

The discovery of the human genetic sequence has been helpful for the identification of genetic alterations related to the appearance of certain cancers (Lynch et al., 2007; Norrild et al., 2007). This is the reason why it is important to keep analyzing the information obtained through the knowledge of the human genetic code. However, the amount of information to be analyzed is so vast that in

Accepted manuscript

ACCEPTED MANUSCRIPT

many occasions a codification or simplification of the information becomes necessary for its later analysis. The development of theoretical 2D graph representations for DNA sequences and proteins has been very important for the analysis and comparison between different sequences. These techniques provide a useful vision of the local and global characteristics of the variations and repetition of nucleotides or amino acids throughout a sequence that is not easily analyzable by other methods. The different methods proposed by Gates (Gates, 1986), Nandy (Nandy, 1994), Leong and Mogenthaler (Leong and Morgenthaler, 1995) consist of drawing up a point that corresponds to a base, moving a positive or negative unit on the ordinate or coordinate axes. Nevertheless, many of 2D graphical representations imply some loss of information due to the overlapping of some bases. Randić *et al.* (Randić et al., 2003) carried out a new 2D representation where there is no loss of information in the transfer of the data of a DNA sequence to its mathematical representation. In analogy to this type of methodologies, there is a type of transformation for protein sequence in a HP Lattice type Complex Networks (Bornberg-Bauer, 1997; Chen and Huang, 2005; Chikenji et al., 2006; Jiang and Zhu, 2005; Li et al., 2002). We can calculate Topological Indices (TIs) for all these classes of networks representations in order to make a numerical description of DNA and protein sequences (González-Díaz et al., 2008). These descriptors (also known as connectivity indices) allow establishing a relation between the biological properties in small and large molecules (QSAR/QSPR), like proteins and genes, and their molecular structure; without to rely upon sequence alignment (Caballero et al., 2007; Cai and Chou, 2005; Cruz-Monteagudo et al., 2007; Chou and Cai, 2003; Chou and Cai, 2005; Chou and Shen, 2008a; Estrada et al., 2006; Estrada et al., 2002; Fernandez et al., 2007a; Fernandez et al., 2007b; González-Díaz et al., 2007a; González-Díaz et al., 2007b; González-Díaz and Uriarte, 2005; Hall et al., 2003; Molina et al., 2004; Prado-Prado et al., 2007; Shen and Chou, 2009; Vilar et al., 2006; Xiao et al., 2009a). Therefore, these methodologies could be an alternative to sequence alignment for the study of proteins and genes (Durand et al.,

ACCEPTED MANUSCRIPT

1997; Hansen et al., 1996; Hofacker et al., 2002; Lecompte et al., 2001; Persson, 2000; Standley et al., 2001; Zhang and Madden, 1997). QSAR methodology has been applied successfully in different situations, with small molecules and proteins, which validate the goodness of the method used in this paper. Partial Least Square (PLS) was used to establish a relationship between Heuristic Molecular Lipophilicity Potential indices and the bioactivity of pyrazole derivatives (Du et al., 2005). New QSAR methods were proposed, such as Multiple Field 3D-QSAR (a combination of classical 2D-QSAR and 3D-QSAR) and Fragment-Based QSAR applied to neuraminidase inhibitors (Du et al., 2008a; Du et al., 2009). We also have different examples of QSAR with proteins with good results in the literature (Du et al., 2008b).

In this work we start from the sequence of different proteins codified by genes implied in cancerous processes, as the Colorectal Cancer, and we transformed these sequences into representations of HP-Lattice type Network. Once the networks are generated, we used MARCH-INSIDE software in order to calculate a series of electrostatic potentials (ξ_k) of the protein sequences forced to fold in the 2D HP Lattice Network. Subsequently, through a Linear Discriminant Analysis (LDA) we found a QSAR model that allows us to discriminate with a high percentage of accuracy between HCCp and no-HCCp. This methodology could be applied to predict Human Colorectal Cancer biomarkers and to understand much better the biological aspects of this disease.

Accepted manuscript

2. Materials and Methods

2.1. *Database*

The used database had been described previously by Sjöblom *et al*. (Sjöblom et al., 2006). All the reported genes related to the Colorectal Cancer (69 genes) with their respective protein sequences had been compiled from this database. The group control is formed by 200 proteins non-related to the appearance of this type of cancer. These corresponding no-HCCp sequence have been used and

ACCEPTED MANUSCRIPT

filtered in addition as examples of useful human proteins in other QSAR studies by Dobson (Dobson and Doig, 2003; Dobson and Doig, 2005). Therefore, 200 non-carcinogenic proteins had been evaluated in order to validate the obtained theoretical model (Dobson and Doig, 2003; Dobson and Doig, 2005).

2.2. *HP Lattice Network Representation*

The different protein sequences of this study were introduced in the software MARCH-INSIDE 2.0 (González-Díaz et al., 2005a). Every sequence was represented through a HP-Lattice type Network. We have been following a recently introduced methodology, analogous to the DNA representations of Nandy, but adapted to the protein study (Aguero-Chapin et al., 2006). Depending on the characteristics of the amino acids, the method groups the 20 types of amino acids in four groups: polar, non-polar, acid or basic amino acids. A HP-Lattice type Network for a determined protein sequence can be seen in Table 1. A characteristic of this type of representations is that the number of nodes (n) in the HP Lattice Network can be equal or smaller than the number of amino acids in the protein. The acid-bases classification prevails over the polar/non-polar and the different groups do not overlap each other. Subsequently, each amino acid in the sequence is placed in a 2D Lattice defined by a Cartesian space with center at the (0, 0) coordinates. The coordinates of the successive bases are calculated to form a Lattice Network with step equal to 1, as follows:

a) Increases in +1 the abscissas axis coordinate for an acid amino acid (rightwards-step) or:

b) Decreases in -1 the abscissas axis coordinate for a basic amino acid (leftwards-step) or:

c) Increases in +1 the ordinates axis coordinate for a polar amino acid (upwards-step):

d) Decreases in -1 the ordinates axis coordinate for a non-polar amino acid (downwards-step).

Table 1 comes about here

2.3. *Calculation of the Electrostatic Potentials in the 2D-HP Lattice Network*

Accepted manuscript

ACCEPTED MANUSCRIPT

A series of the Electrostatic Potential (ξ_k) approximations for each 2D-HP Lattice Network had been calculated using the previously mentioned software MARCH-INSIDE 2.0. In order to calculate the different values of ξ_k, the potentials of each amino acid were averaged, considering the values of all nearby amino acids, located to a topological distance not greater than k within the HP Lattice Network. The Markov Chain theory is used to estimate the first six values of ξ_k of every sequence (k = 0, 1, 2, 3, 4, and 5) (González-Díaz et al., 2007c; Saiz-Urra et al., 2005).

The method uses essentially three matrix magnitudes:

a) The matrix $^1\mathbf{\Pi}$ (see Eq. 1). This matrix is built up as a squared matrix (n × n). The matrix $^1\mathbf{\Pi}$ contains the probabilities $^1p_{ij}$ to reach a node n_i with charge q_i moving throughout a walk of length k = 1 from other node n_j with charge qj. We can also speak about the probabilities of reaching a node with electrostatic potential φ_i moving from another node:

$$^1p_{ij} = \frac{\alpha_{ij} \cdot \left(\frac{q_j}{r^2_{0,j}} \right)}{\sum_{m=1}^{n} \alpha_{im} \cdot \left(\frac{q_m}{r^2_{0,m}} \right)} = \frac{\alpha_{ij} \cdot \varphi_j}{\sum_{m=1}^{n} \alpha_{im} \cdot \varphi_m} \quad (1) \qquad ^Ap_0(j) = \frac{q_j}{\sum_{m=1}^{n} q_m} \quad (2)$$

b) The charges vector $^0\mathbf{Q}$. The method considers anyhow that a total charge or weight (q_i) can be assigned to each node. This charge of the node is equal to the sum of the charges of all the aminoacids coinciding in the same node. So, in order to retain a more compact matrix notation all charges are arranged as a column vector $^0\mathbf{Q}$.

c) The zero order electrostatic potential vector $^0\boldsymbol{\zeta}$ (see Eq. 2). This vector lists the absolute initial probabilities $^Ap_k(j)$ with which a node selected at random presents a given charge q_j.

Thus, the use of Markov Chains theory allows calculating the average Electrostatic Potential (ξ_k) for all nodes n_j that can be reached in the 2D-HP Lattice Network moving from any node n_i using walks of length k. Considering that the ξ_k are average values associated to a discrete lattice we determine

Accepted manuscript

ACCEPTED MANUSCRIPT

them as the sum of two-terms products using Chapman-Kolmogorov type equations. The readers can see previous works explaining in detail the use of this kind of approach in many different situations (Cruz-Monteagudo et al., 2007; González-Díaz et al., 2003; Gonzalez-Diaz et al., 2007a; González-Díaz et al., 2007d; González-Díaz et al., 2007e; Perez-Bello et al., 2009; Ramos de Armas et al., 2004; Santana et al., 2006). The first term is the probability of arising the node n_j moving from any node n_i throughout walks of length k and the second one the charge of the node q_j (see central member of Eq. 3 below):

$$\xi_k = \sum_{j=1}^{n} {}^{A}p_k(j) \cdot q_j = {}^{0}\boldsymbol{\zeta}^{\mathrm{T}} \cdot \left({}^{1}\boldsymbol{\Pi}\right)^{k} \cdot {}^{0}\mathbf{Q} \qquad (3)$$

Note that the higher-order ξ_k depends only on the absolute probabilities ${}^{A}p_0(j)$ of order 0, on the charges and on the matrix. In particular, the evaluation of such expansions for k = 0 gives the order zero Electrostatic potential (ξ_0); for k = 1 the short-range Electrostatic potential (ξ_1), for k = 2 the middle-range Electrostatic potential (ξ_2), and for k > 2 the long-range Electrostatic potentials ($\xi_{k>2}$). This expansion is illustrated for the linear graph n_1-n_2-n_3 characteristic of the sequence (Asp-Glu-Asp-Lys), please note that the central node contains both Glu and Asp: (Gonzalez-Diaz et al., 2005b)

$$\xi_0 = \left[{}^{A}p_0(n1),\ {}^{A}p_0(n2),\ {}^{A}p_0(n3)\right] \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} = {}^{A}p_0(n1) \cdot q_1 + {}^{A}p_0(n2) \cdot q_2 + {}^{A}p_0(n3) \cdot q_3 \qquad (3a)$$

$$\xi_1 = \left[{}^{A}p_0(n1),\ {}^{A}p_0(n2),\ {}^{A}p_0(n3)\right] \cdot \begin{bmatrix} {}^{1}p_{11} & {}^{1}p_{12} & 0 \\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23} \\ 0 & {}^{1}p_{32} & {}^{1}p_{33} \end{bmatrix} \cdot \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} = {}^{A}p_1(n1) \cdot q_1 + {}^{A}p_1(n2) \cdot q_2 + {}^{A}p_1(n3) \cdot q_3 \qquad (3b)$$

$$\xi_2 = \left[{}^{A}p_0(n1),\ {}^{A}p_0(n2),\ {}^{A}p_0(n3)\right] \cdot \begin{bmatrix} {}^{1}p_{11} & {}^{1}p_{12} & 0 \\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23} \\ 0 & {}^{1}p_{32} & {}^{1}p_{33} \end{bmatrix} \cdot \begin{bmatrix} {}^{1}p_{11} & {}^{1}p_{12} & 0 \\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23} \\ 0 & {}^{1}p_{32} & {}^{1}p_{33} \end{bmatrix} \cdot \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \qquad (3c)$$

$$\xi_k = \left[{}^{A}p_0(n1),\ {}^{A}p_0(n2),\ {}^{A}p_0(n3)\right] \cdot \begin{bmatrix} {}^{1}p_{11} & {}^{1}p_{12} & 0 \\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23} \\ 0 & {}^{1}p_{32} & {}^{1}p_{33} \end{bmatrix} \cdot \begin{bmatrix} {}^{1}p_{11} & {}^{1}p_{12} & 0 \\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23} \\ 0 & {}^{1}p_{32} & {}^{1}p_{33} \end{bmatrix} \cdots\cdots \left(\begin{bmatrix} {}^{1}p_{11} & {}^{1}p_{12} & 0 \\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23} \\ 0 & {}^{1}p_{32} & {}^{1}p_{33} \end{bmatrix} \right)_{k-2} \cdot \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \qquad (3d)$$

2.4. *Statistical Analysis*

Accepted manuscript

ACCEPTED MANUSCRIPT

At this level a Linear Discriminant Analysis (Hill and Lewicki, 2006 ; Zhu and Martinez, 2006) has been carried out relating the different descriptors previously calculated to the carcinogenic activity of the different protein sequences. The Software STATISTICA 6.0 (StatSoft.Inc., 2002) package has been used for the statistical analysis and to develop a function of classification that differentiates between HCCp and no-HCCp:

$$HCC - score = b_0 \cdot \xi_0 + b_1 \cdot \xi_1 ... + b_k \cdot \xi_k - a_0 \qquad (4)$$

The variable HCC-score is the biological property under investigation, in this case the Human Colorectal Cancer (HCC), ξ_k are electrostatic potential descriptors calculated for the database and b_k and a_0 are the coefficients obtained by the Linear Discriminant Analysis (LDA). The function was obtained by using the forward-stepwise method for a variable selection. In the development of this classification function the values of 1 and 0 were assigned to HCCp and no-HCCp respectively. The statistical parameters that define the quality of the model are the Wilks' statistic (*U*), Fisher ratio (*F*) and the percentage of good classification for the training and CV sets. The *a posteriori* probability calculated from the Mahalanobis distance was used for the classification of cases as active or inactive (Hill and Lewicki, 2006).

Accepted manuscript

3. Results and Discussion

A superposition of all the HP Lattice Network has been done, including the HCCp and no-HCCp (see Figure 1). This superposition shows that there is no clear differentiation between HCCp and no-HCCp networks. This is the reason why the calculation of TIs in combination with the discriminant analysis is necessary.

Figure 1 comes about here

ACCEPTED MANUSCRIPT

The linear classification model derived from the STATISTICA package is given below together with the statistical parameters:

$$HCC - score = 39.736 \times \xi_{02} - 37.057 \times \xi_{03} - 2.197 \qquad (5)$$

$$N = 269 \qquad U = 0.528 \qquad F\,(2,266) = 118.926 \qquad p<0.05$$

Here, N is the number of compounds included in the discriminant analysis calculation, U is the Wilks' statistics, F is the Fisher ratio and p is the significance level. The proteins that have been introduced in the model and their *a posteriori* probabilities calculated in the analysis are shown in Table 2. The described model presents a good classification percentage of 89.86 % in the positive cases (62 of 69 proteins codified by carcinogenic genes are recognized by the model). The negative cases are evaluated by an 83.00 % of accuracy (166 of 200 proteins non-implied in the cancerous processes are well evaluated).

Table 2 comes about here

The parameters refer to the model with all the cases in the training. But it is worth noting that to find the model we have divided the database into two series: training and CV (cross-validation) series. In order to demonstrate the stability of the model this process has been repeated on different occasions randomly interchanging cases in both series. The obtained results are similar to those of the previously described model from the quality of the statistical parameters point of view as in the percentage of good classification (see Table 3). In statistical prediction, three cross-validation methods are often used: subsampling test, independent dataset test, and jackknife test (Chou and Zhang, 1995). However, as demonstrated in (Chou and Shen, 2007), the jackknife test has the least arbitrariness and therefore has been increasingly and widely used to test various prediction methods (see, e.g., (Chen et al., 2008; Chen and Han, 2009; Chou and Shen, 2008a; Chou and Shen, 2008b; Chou and Shen, 2009; Ding et al., 2009a; Ding et al., 2009b; Du and Li, 2008; Georgiou et al., 2009;

Accepted manuscript

ACCEPTED MANUSCRIPT

Kannan et al., 2008; Li and Li, 2008; Lin, 2008; Lin et al., 2008; Lin et al., 2009; Munteanu et al., 2008; Nanni and Lumini, 2009; Rezaei et al., 2008; Shen and Chou, 2009; Shen et al., 2009; Shi et al., 2008; Tian et al., 2008; Wang et al., 2008; Xiao et al., 2009b; Zeng et al., 2009; Zhang and Fang, 2008; Zhang et al., 2008). In the jackknife or leave-one-out test each case in the database is predicted for the model constructed using all the cases except the one being predicted. In this paper, we used the three methods. As we described above, the model was found dividing the database in training and CV series (subsampling test). A jackknife validation procedure was carried out in the Table 2. The results in the good classification percentage were maintained when we applied this validation method. Another method for validating this model consists of the evaluation of 200 proteins that are not related to the Colorectal Cancer and, therefore, must be evaluated by our model as non-active (see Table A of the Supporting Information). The percentage of good classification in this series of external prediction is 74% (148 of these 20 proteins are classified correctly)

Table 3 comes about here

Receiver Operating Characteristic (ROC) curve for the model was also calculated (Diamond, 1987; Hanley, 1989; Mann et al., 1992; Metz et al., 1973). The fraction of true positives (Sensitivity) is contrasted with the fraction of false positives (1-Specificity) in this type of curves. The area under the curve can take values between 1 (perfect classifier) and 0.5 (useless random classifier). In our ROC curve the area under the curve is 0.96, which confirms that the used model is not a random classifier (see Figure 2).

Figure 2 comes about here

We have made also a leverage-based analysis of the domain of applicability (DA) of the model (see Figure 3) (Hill and Lewicki, 2006 ; Merli, 2005). Through this type of studies we can see the applicability limits of our methodology. This way we can avoid making erroneous predictions for some cases that would be out of the model DA. A Cartesian double ordinate, where the CV residuals

Accepted manuscript

ACCEPTED MANUSCRIPT

and standard residuals are represented, is shown in the Figure 3. The leverage is represented on the abscissas axis. The DA of the model is defined as a squared area within ±2 band for residuals and a leverage threshold of h = 0.033.

Figures 3 comes about here

4. Conclusions

The combination of the 2D representations of the protein sequence, as the HP Lattice Network, and the calculation of TIs constitute a significant tool for designing theoretical models that relate the protein structure to its biological properties. This type of methodologies can be complementary to the methods of sequence alignment for studying protein databases. The approximation we described in this work could represent a method that makes possible the analysis of large protein databases in order to identify novel proteins susceptible of the development of colorectal cancer and to understand better the biological phenomena related to cancer.

Supporting Information Available. The external prediction series for the theoretical model with 200 proteins is shown in Table A of the Supporting Information.

Accepted manuscript

Acknowledgements

We are grateful to the Xunta de Galicia (PGIDIT05PXIB20304PR, PGIDIT05BTF20302PR and INCITE07PXI203039ES) and Ministerio de Sanidad y Consumo (FISPI061457) for financial support. We thank the reviewers for their constructive comments that led us to craft an improved manuscript.

ACCEPTED MANUSCRIPT

References

Aguero-Chapin, G., Gonzalez-Diaz, H., Molina, R., Varona-Santos, J., Uriarte, E., and Gonzalez-Diaz, Y., 2006. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L. FEBS Lett 580, 723-730.

Bornberg-Bauer, E., 1997. How are model protein structures distributed in sequence space? Biophys J 73, 2393-2403.

Boursi, B., and Arber, N., 2007. Current and future clinical strategies in colon cancer prevention and the emerging role of chemoprevention. Curr Pharm Des 13, 2274-2282.

Caballero, J., Fernandez, L., Garriga, M., Abreu, J.I., Collina, S., and Fernandez, M., 2007. Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines. J Mol Graph Model 26, 166-178.

Cai, Y.D., and Chou, K.C., 2005. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J Proteome Res 4, 967-971.

Cruz-Monteagudo, M., González-Díaz, H., Agüero-Chapin, G., Santana, L., Borges, F., Domínguez, R.E., Podda, G., and Uriarte, E., 2007. Computational Chemistry Development of a Unified Free Energy Markov Model for the Distribution of 1300 Chemicals to 38 Different Environmental or Biological Systems. J Comput Chem 28, 1909-1922.

Chen, C., Chen, L.X., Zou, X.Y., and Cai, P.X., 2008. Predicting protein structural class based on multi-features fusion. J Theor Biol 253, 388-392.

Chen, M., and Huang, W.Q., 2005. A branch and bound algorithm for the protein folding problem in the HP lattice model. Genomics Proteomics Bioinformatics 3, 225-230.

Chen, Y., and Han, K., 2009. BSFINDER: Finding Binding Sites of HCV Proteins Using a Support Vector Machine. Protein Peptide Lett 16, 373-382.

Chikenji, G., Fujitsuka, Y., and Takada, S., 2006. Shaping up the protein folding funnel by local interaction: lesson from a structure prediction study. Proc Natl Acad Sci U S A 103, 3141-6.

Chou, K.C., and Cai, Y.D., 2003. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. J Cell Biochem 90, 1250-1260.

Chou, K.C., and Cai, Y.D., 2005. Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inf Model 45, 407-413.

Chou, K.C., and Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. Anal Biochem 370, 1-16.

Chou, K.C., and Shen, H.B., 2008a. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem Biophys Res Commun 376, 321-325.

Chou, K.C., and Shen, H.B., 2008b. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3, 153-162.

Chou, K.C., and Shen, H.B., 2009. FoldRate: A web-server for predicting protein folding rates from primary sequence. The Open Bioinformatics Journal 3, 31-50.

Chou, K.C., and Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30, 275-349.

Diamond, G.A., 1987. ROC steady: a receiver operating characteristic curve that is invariant relative to selection bias. Med Decis Making 7, 238-243.

Ding, H., Luo, L., and Lin, H., 2009a. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein Peptide Lett 16, 351-355.

Accepted manuscript

ACCEPTED MANUSCRIPT

Ding, Y.S., Zhang, T.L., Gu, Q., Zhao, P.Y., and Chou, K.C., 2009b. Using maximum entropy model to predict protein secondary structure with single sequence. Protein Peptide Lett 16, 552-560.

Dobson, P.D., and Doig, A.J., 2003. Distinguishing enzyme structures from non-enzymes without alignments. J Mol Biol 330, 771-783.

Dobson, P.D., and Doig, A.J., 2005. Predicting enzyme class from protein structure without alignments. J Mol Biol 345, 187-199.

Du, P., and Li, Y., 2008. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. J Theor Biol 253, 579-589.

Du, Q.S., Mezey, P.G., Chou, K.C., 2005. Heuristic Molecular Lipophilicity Potential (HMLP): A 2D-QSAR Study to LADH of Molecular Family Pyrazole and Derivatives. J Comput Chem 26, 461-470.

Du, Q.S., Huang, R.B., Wei, Y.T., Du, L.Q., Chou, K.C., 2008a. Multiple Field Three Dimensional Quantitative Structure-Activity Relationship (MF-3D-QSAR). J Comput Chem 29, 211-219.

Du, Q.S., Huang, R.B., Chou, K.C., 2008b. Review: Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design, Curr Protein Pept Sci 9, 248-259.

Du, Q.S., Huang, R.B., Wei, Y.T., Pang, Z.W., Du, L.Q., Chou, K.C., 2009. Fragment-Based Quantitative Structure-Activity Relationship (FB-QSAR) for Fragment-Based Drug Design, J Comput Chem 30, 295-304.

Durand, P., Canard, L., and Mornon, J.P., 1997. Visual BLAST and visual FASTA: graphic workbenches for interactive analysis of full BLAST and FASTA outputs under MICROSOFT WINDOWS 95/NT. Comput Appl Biosci 13, 407-413.

Estrada, E., Uriarte, E., and Vilar, S., 2006. Effect of Protein Backbone Folding on the Stability of Protein-Ligand Complexes. J Proteome Res 5, 105-111.

Estrada, E., Vilar, S., Uriarte, E., and Gutierrez, Y., 2002. In silico studies toward the discovery of new anti-HIV nucleoside compounds with the use of TOPS-MODE and 2D/3D connectivity indices. 1. Pyrimidyl derivatives. J Chem Inf Comput Sci 42, 1194-1203.

Fernandez, M., Caballero, J., Fernandez, L., Abreu, J.I., and Garriga, M., 2007a. Protein radial distribution function (P-RDF) and Bayesian-Regularized Genetic Neural Networks for modeling protein conformational stability: chymotrypsin inhibitor 2 mutants. J Mol Graph Model 26, 748-759.

Fernandez, M., Caballero, J., Fernandez, L., Abreu, J.I., and Acosta, G., 2007b. Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines. Proteins 70, 167-175.

Gates, M.A., 1986. A simple way to look at DNA. J Theor Biol 119, 319-328.

Georgiou, D.N., Karakasidis, T.E., Nieto, J.J., and Torres, A., 2009. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J Theor Biol 257, 17-26.

González-Díaz, H., de Armas, R.R., and Molina, R., 2003. Markovian negentropies in bioinformatics. 1. A picture of footprints after the interaction of the HIV-1 Psi-RNA packaging region with drugs. Bioinformatics 19, 2079-2087.

González-Díaz, H., Molina-Ruiz, R., and Hernandez, I., MARCH-INSIDE version 2.0 (Markovian Chemicals In Silico Design), 2005a, pp. MARCH-INSIDE version 2.0 (Markovian Chemicals In Silico Design). Main author information requesting contact email: gonzalezdiazh@yahoo.es.

Gonzalez-Diaz, H., Molina, R., and Uriarte, E., 2005b. Recognition of stable protein mutants with 3D stochastic average electrostatic potentials. FEBS Lett 579, 4297-4301.

Accepted manuscript

ACCEPTED MANUSCRIPT

González-Díaz, H., and Uriarte, E., 2005. Proteins QSAR with Markov average electrostatic potentials. Bioorg Med Chem Lett 15, 5088-5094.

González-Díaz, H., Agüero-Chapin, G., Varona, J., Molina, R., Delogu, G., Santana, L., Uriarte, E., and Gianni, P., 2007a. 2D-RNA-Coupling Numbers: A New Computational Chemistry Approach to Link Secondary Structure Topology with Biological Function. J Comput Chem 28, 1049–1056.

Gonzalez-Diaz, H., Vilar, S., Santana, L., Podda, G., and Uriarte, E., 2007b. On the applicability of QSAR for recognition of miRNA bioorganic structures at early stages of organism and cell development: embryo and stem cells. Bioorg Med Chem 15, 2544-2550.

González-Díaz, H., Pérez-Castillo, Y., Podda, G., and Uriarte, E., 2007c. Computational Chemistry Comparison of Stable/Nonstable Protein Mutants Classification Models Based on 3D and Topological Indices. J Comput Chem 28, 1990-1995.

González-Díaz, H., Vilar, S., Santana, L., and Uriarte, E., 2007d. Medicinal Chemistry and Bioinformatics – Current Trends in Drugs Discovery with Networks Topological Indices. Curr Top Med Chem 7, 1025-1039.

González-Díaz, H., Saiz-Urra, L., Molina, R., Santana, L., and Uriarte, E., 2007e. A Model for the Recognition of Protein Kinases Based on the Entropy of 3D van der Waals Interactions. J Proteome Res 6, 904-908.

González-Díaz, H., González-Díaz, Y., Santana, L., Martínez-Ubeira, F., and Uriarte, E., 2008. Proteomics, networks and connectivity indices. Proteomics 8, 750-778.

Hall, L.M., Hall, L.H., and Kier, L.B., 2003. Modeling drug albumin binding affinity with e-state topological structure representation. J Chem Inf Comput Sci 43, 2120-2128.

Hanley, J.A., 1989. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging 29, 307-335.

Hansen, J.E., Lund, O., Nielsen, J.O., Brunak, S., and Hansen, J.E., 1996. Prediction of the secondary structure of HIV-1 gp120. Proteins 25, 1-11.

Hill, T., and Lewicki, P., 2006 STATISTICS Methods and Applications. StatSoft, Tulsa.

Hofacker, I.L., Fekete, M., and Stadler, P.F., 2002. Secondary structure prediction for aligned RNA sequences. J Mol Biol 319, 1059-1066.

Jiang, M., and Zhu, B., 2005. Protein folding on the hexagonal lattice in the HP model. J Bioinform Comput Biol 3, 19-34.

Kannan, S., Hauth, A.M., and Burger, G., 2008. Function prediction of hypothetical proteins without sequence similarity to proteins of known function. Protein Peptide Lett 15, 1107-1116.

Lecompte, O., Thompson, J.D., Plewniak, F., Thierry, J., and Poch, O., 2001. Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene 270, 17-30.

Leong, P.M., and Morgenthaler, S., 1995. Random walk and gap plots of DNA sequences. Comput Appl Biosci 11, 503-507.

Li, F.M., and Li, Q.Z., 2008. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Peptide Lett 15, 612-616.

Li, H., Tang, C., and Wingreen, N.S., 2002. Designability of protein structures: a lattice-model study using the Miyazawa-Jernigan matrix. Proteins 49, 403-412.

Lin, H., 2008. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252, 350-356.

Lin, H., Ding, H., Feng-Biao Guo, F.B., Zhang, A.Y., and Huang, J., 2008. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Peptide Lett 15, 739-744.

Accepted manuscript

ACCEPTED MANUSCRIPT

Lin, Z.H., Wang, H.L., Zhu, B., Wang, Y.Q., Lin, Y., and Wu, Y.Z., 2009. Estimation of Affinity of HLA-A*0201 Restricted CTL Epitope Based on the SCORE Function. Protein Peptide Lett 16, 561-569.

Luchtenborg, M., White, K.K., Wilkens, L., Kolonel, L.N., and Le Marchand, L., 2007. Smoking and colorectal cancer: different effects by type of cigarettes? Cancer Epidemiol Biomarkers Prev 16, 1341-1347.

Lynch, H.T., Fusaro, R.M., and Lynch, J.F., 2007. Hereditary cancer syndrome diagnosis: molecular genetic clues and cancer control. Future Oncol 3, 169-181.

Mann, F.A., Hildebolt, C.F., and Wilson, A.J., 1992. Statistical analysis with receiver operating characteristic curves. Radiology 184, 37-38.

Merli, M., 2005. Outlier recognition in crystal-structure least-squares modelling by diagnostic techniques based on leverage analysis. Acta Crystallogr A 61, 471-477.

Metz, C.E., Goodenough, D.J., and Rossmann, K., 1973. Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. Radiology 109, 297-303.

Molina, E., González-Díaz, H., Gonzalez, M.P., Rodriguez, E., and Uriarte, E., 2004. Designing antibacterial compounds through a topological substructural approach. J Chem Inf Comput Sci 44, 515-521.

Munteanu, C.B., Gonzalez-Diaz, H., and Magalhaes, A.L., 2008. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. J Theor Biol 254, 476-482.

Nandy, A., 1994. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Curr Sci 66, 309-314.

Nanni, L., and Lumini, A., 2009. A Further Step Toward an Optimal Ensemble of Classifiers for Peptide Classification, a Case Study: HIV Protease. Protein Peptide Lett 16, 163-167.

Norrild, B., Guldberg, P., and Ralfkiaer, E., 2007. Cancer genomics. Apmis 115, 1037-1038.

Perez-Bello, A., Munteanu, C.R., Ubeira, F.M., Lopes De Magalhaes, A., Uriarte, E., and González-Díaz, H., 2009. Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices. J Theor Biol 256, 458-466.

Persson, B., 2000. Bioinformatics in protein analysis. Exs 88, 215-231.

Prado-Prado, F., González-Díaz, H., Santana, L., and Uriarte, E., 2007. Unified QSAR approach to antimicrobials. Part 2: Predicting activity against more than 90 different species in order to halt antibacterial resistance. Bioorg Med Chem 15 897-902.

Ramos de Armas, R., González-Díaz, H., Molina, R., and Uriarte, E., 2004. Markovian Backbone Negentropies: Molecular descriptors for protein research. I. Predicting protein stability in Arc repressor mutants. Proteins 56, 715-723.

Randić, M., Vračko, M., Lerš, N., and Plavšić, D., 2003. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett 368, 1-6.

Rezaei, M.A., Abdolmaleki, P., Karami, Z., Asadabadi, E.B., Sherafat, M.A., Abrishami-Moghaddam, H., Fadaie, M., and Forouzanfar, M., 2008. Prediction of membrane protein types by means of wavelet analysis and cascaded neural networks. J Theor Biol 254, 817-820.

Saiz-Urra, L., González-Díaz, H., and Uriarte, E., 2005. Proteins Markovian 3D-QSAR with spherically-truncated average electrostatic potentials. Bioorg Med Chem 13, 3641-3647.

Santana, L., Uriarte, E., González-Díaz, H., Zagotto, G., Soto-Otero, R., and Mendez-Alvarez, E., 2006. A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. J Med Chem 49, 1149-1156.

Accepted manuscript

ACCEPTED MANUSCRIPT

Schafmayer, C., Buch, S., Egberts, J.H., Franke, A., Brosch, M., El Sharawy, A., Conring, M., Koschnick, M., Schwiedernoch, S., Katalinic, A., Kremer, B., Folsch, U.R., Krawczak, M., Fandrich, F., Schreiber, S., Tepel, J., and Hampe, J., 2007. Genetic investigation of DNA-repair pathway genes PMS2, MLH1, MSH2, MSH6, MUTYH, OGG1 and MTH1 in sporadic colon cancer. Int J Cancer 121, 555-558.

Shen, H.B., and Chou, K.C., 2009. QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. J Proteome Res 8, 1577-1584.

Shen, H.B., Song, J.N., and Chou, K.C., 2009. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. Journal of Biomedical Science and Engineering (JBiSE) 2, 136-143 (open accessible at http://www.srpublishing.org/journal/jbise/).

Shi, M.G., Huang, D.S., and Li, X.L., 2008. A Protein Interaction Network Analysis for Yeast Integral Membrane Protein. Protein Peptide Lett 15, 692-699.

Sjöblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S.D., Willis, J., Dawson, D., Willson, J.K., Gazdar, A.F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B.H., Bachman, K.E., Papadopoulos, N., Vogelstein, B., Kinzler, K.W., and Velculescu, V.E., 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314, 268-274.

Standley, D.M., Eyrich, V.A., An, Y., Pincus, D.L., Gunn, J.R., and Friesner, R.A., 2001. Protein structure prediction using a combination of sequence-based alignment, constrained energy minimization, and structural alignment. Proteins Suppl 5, 133-139.

StatSoft.Inc., STATISTICA (data analysis software system), version 6.0, www.statsoft.com.Statsoft, Inc., Tulsa, OK, U.S.A. 2002, pp. STATISTICA (data analysis software system), version 6.0, www.statsoft.com.Statsoft.

Stein, U., and Schlag, P.M., 2007. Clinical, biological, and molecular aspects of metastasis in colorectal cancer. Recent Results Cancer Res 176, 61-80.

Tian, F., Lv, F., Zhou, P., Yang, Q., and Jalbout, A.F., 2008. Toward prediction of binding affinities between the MHC protein and its peptide ligands using quantitative structure-activity relationship approach. Protein Peptide Lett 15, 1033-1043.

Vilar, S., Santana, L., and Uriarte, E., 2006. Probabilistic neural network model for the in silico evaluation of anti-HIV activity and mechanism of action. J Med Chem 49, 1118-1124.

Wang, T., Yang, J., Shen, H.B., and Chou, K.C., 2008. Predicting membrane protein types by the LLDA algorithm. Protein Peptide Lett 15, 915-921.

Xiao, X., Wang, P., Chou, K.C., 2009a. Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. J Appl Crystallogr 42, 169-173.

Xiao, X., Wang, P., and Chou, K.C., 2009b. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J Comput Chem 30, 1414-1423.

Young, G.P., 2007. Diet and genomic stability. Forum Nutr 60, 91-96.

Zeng, Y.H., Guo, Y.Z., Xiao, R.Q., Yang, L., Yu, L.Z., and Li, M.L., 2009. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J Theor Biol 259, 366-372.

Zhang, G.Y., and Fang, B.S., 2008. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. J Theor Biol 253, 310-315.

Accepted manuscript

ACCEPTED MANUSCRIPT

Zhang, G.Y., Li, H.C., and Fang, B.S., 2008. Predicting lipase types by improved Chou's pseudo-amino acid composition. Protein Peptide Lett 15, 1132-1137.

Zhang, J., and Madden, T.L., 1997. PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. Genome Res 7, 649-656.

Zhu, M., and Martinez, A.M., 2006. Subclass discriminant analysis. IEEE Trans Pattern Anal Mach Intell 28, 1274-1286.

Accepted manuscript

ACCEPTED MANUSCRIPT

Figure and table captions

Table 1. Description of Nodes, Aminoacid Sequence, Coordinates and Stochastic Matrix for a HP Lattice Network.

Table 2. Proteins Introduced in the Model and their *a Posteriori* Probabilities (P). 62 of 69 Positive Cases are well Evaluated (89.86% of Good Classification). 166 of 200 Negative Cases are well Evaluated (83.00% of Good Classification). P (jackknife) is the *a Posteriori* Probability Extracted from the Jackknife Test.

Table 3. Description of the Models Varying the Cross-validation Series.

Table A (Supporting Information). External prediction series: 148 of 200 proteins non-related to Colorectal Cancer are correctly classified (74.0% of good classification). P represents the a posteriori probability of not having a carcinogenic activity.

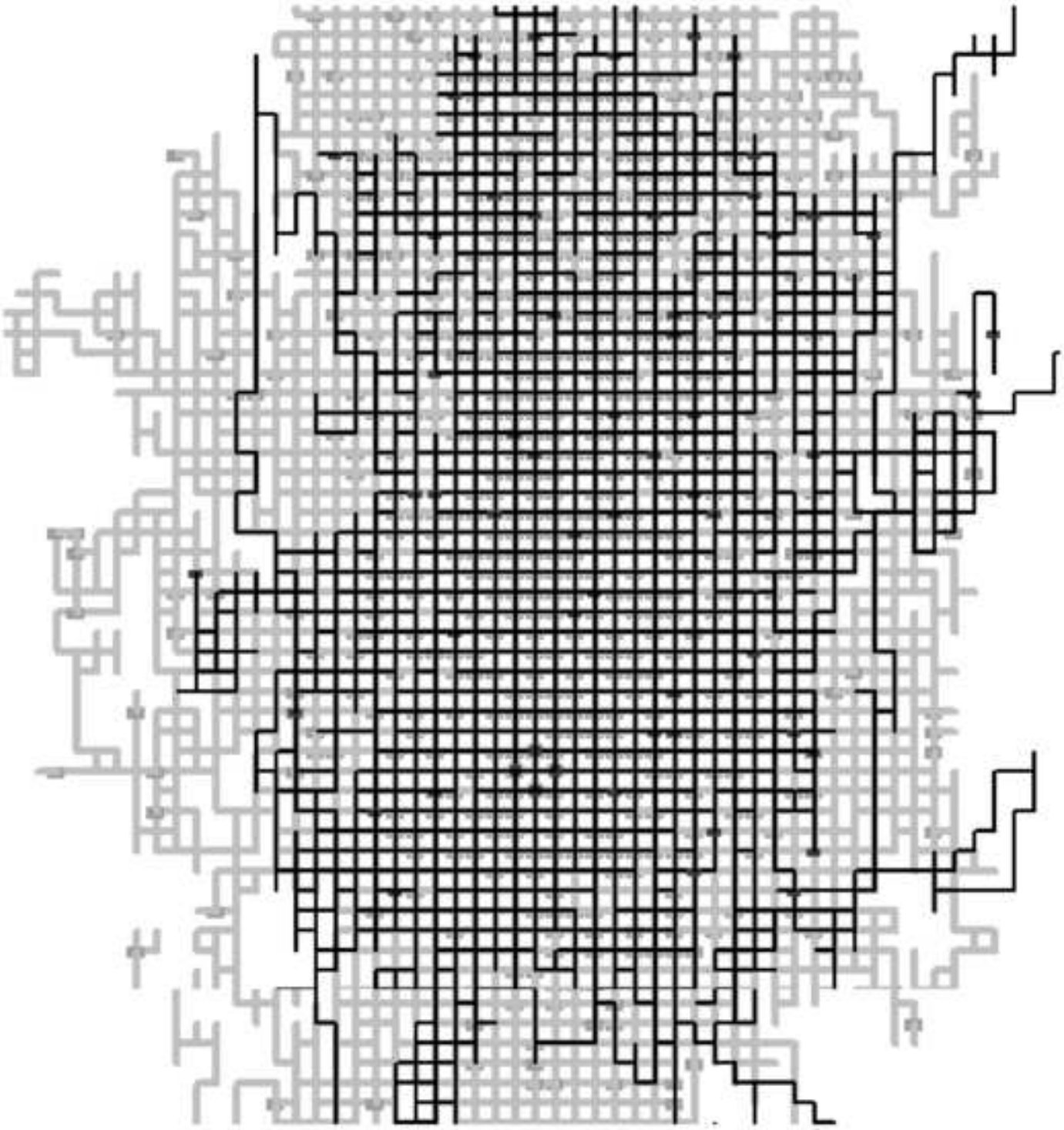
Figure 1. Superposition of all HP Lattice Network, including the HCCp (black coloured) and non-HCCp (grey coloured).

Figure 2. Receiver Operating Characteristic (ROC) curve for the model. The area under the curve is 0.96.

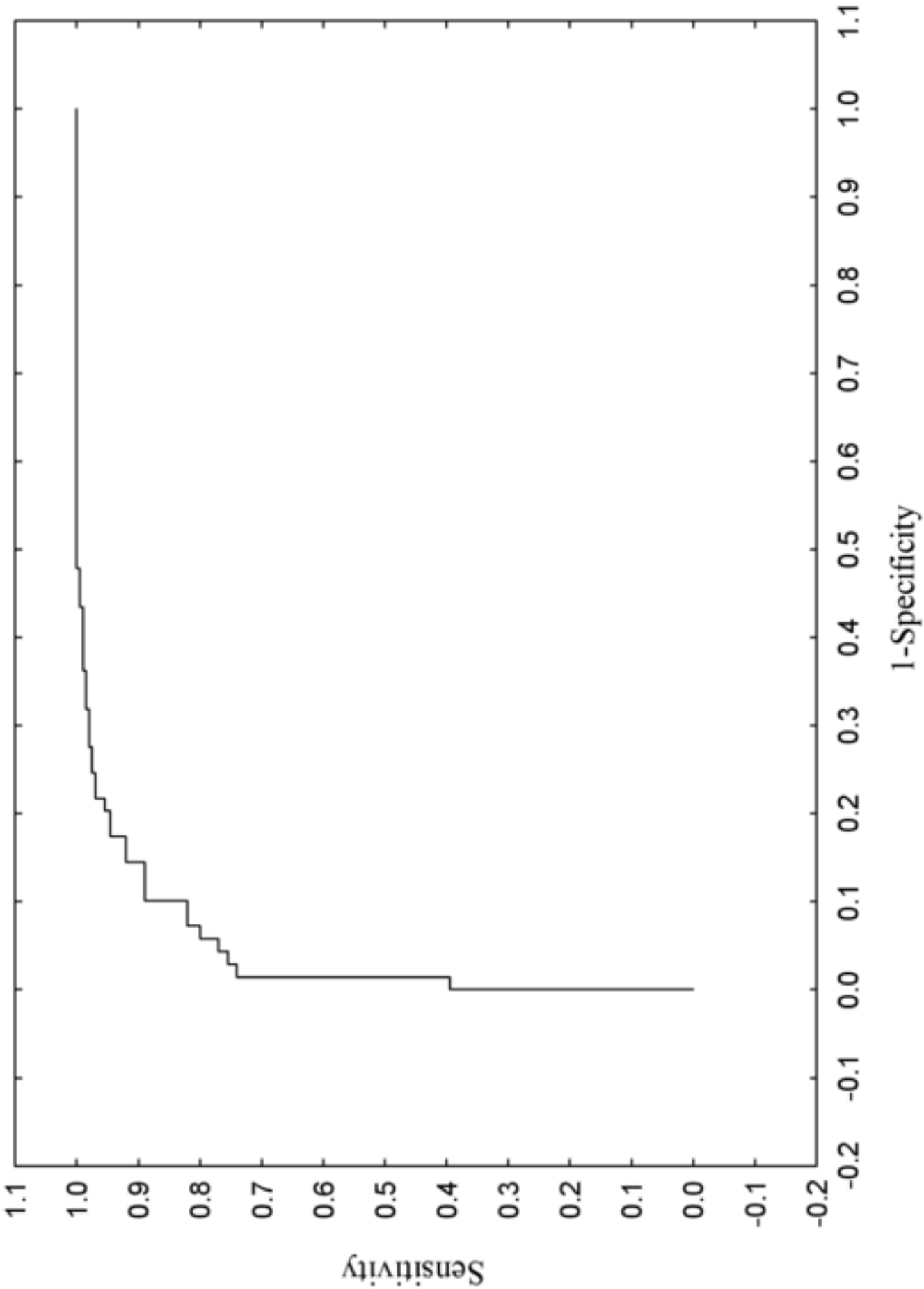
Figure 3. Leverage-based analysis of the domain of applicability of the model: a) Colorectal Cancer Proteins, b) No Colorectal Cancer Proteins; model leverage threshold is $h = 0.033$.

Accepted manuscript

ACCEPTED MANUSCRIPT

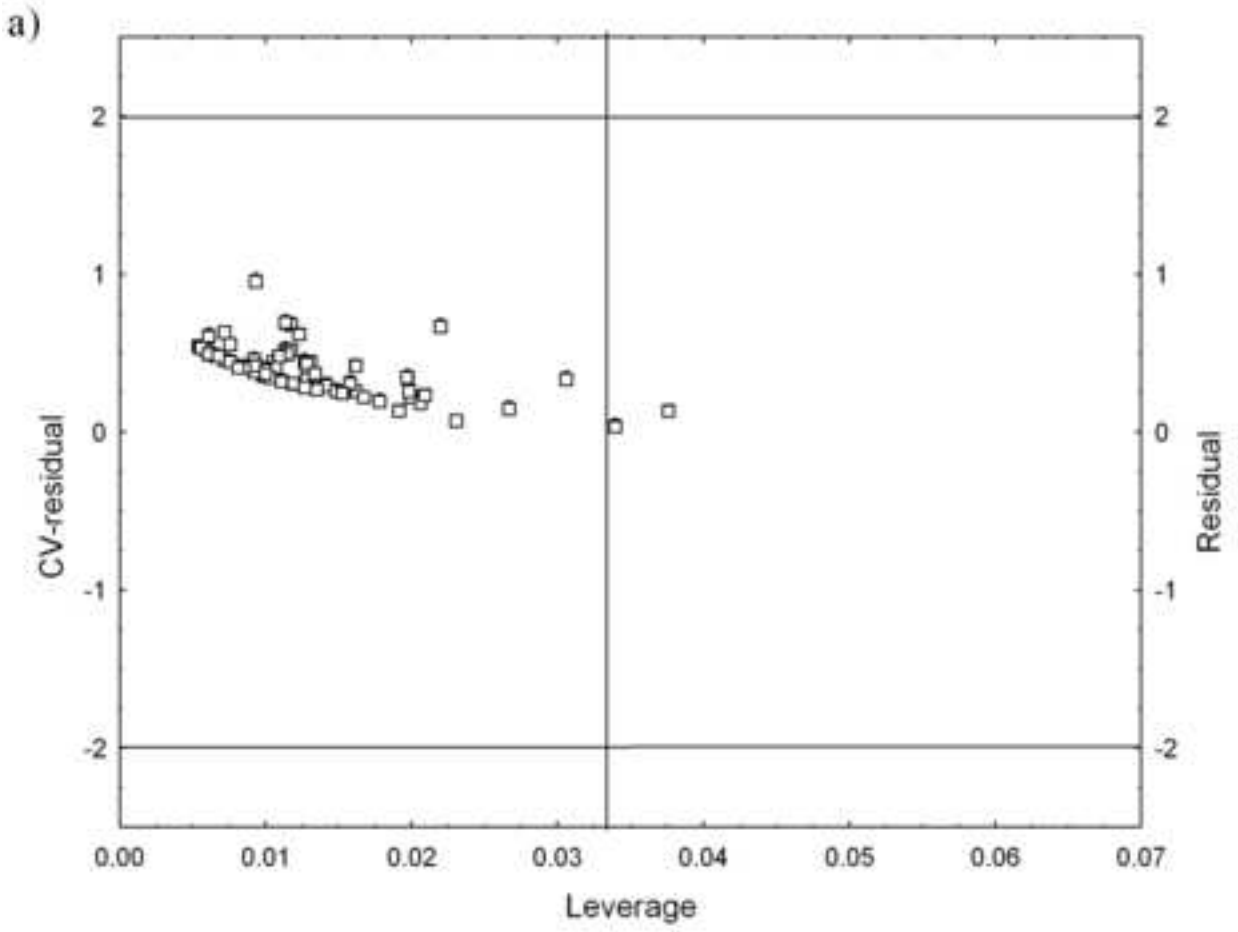


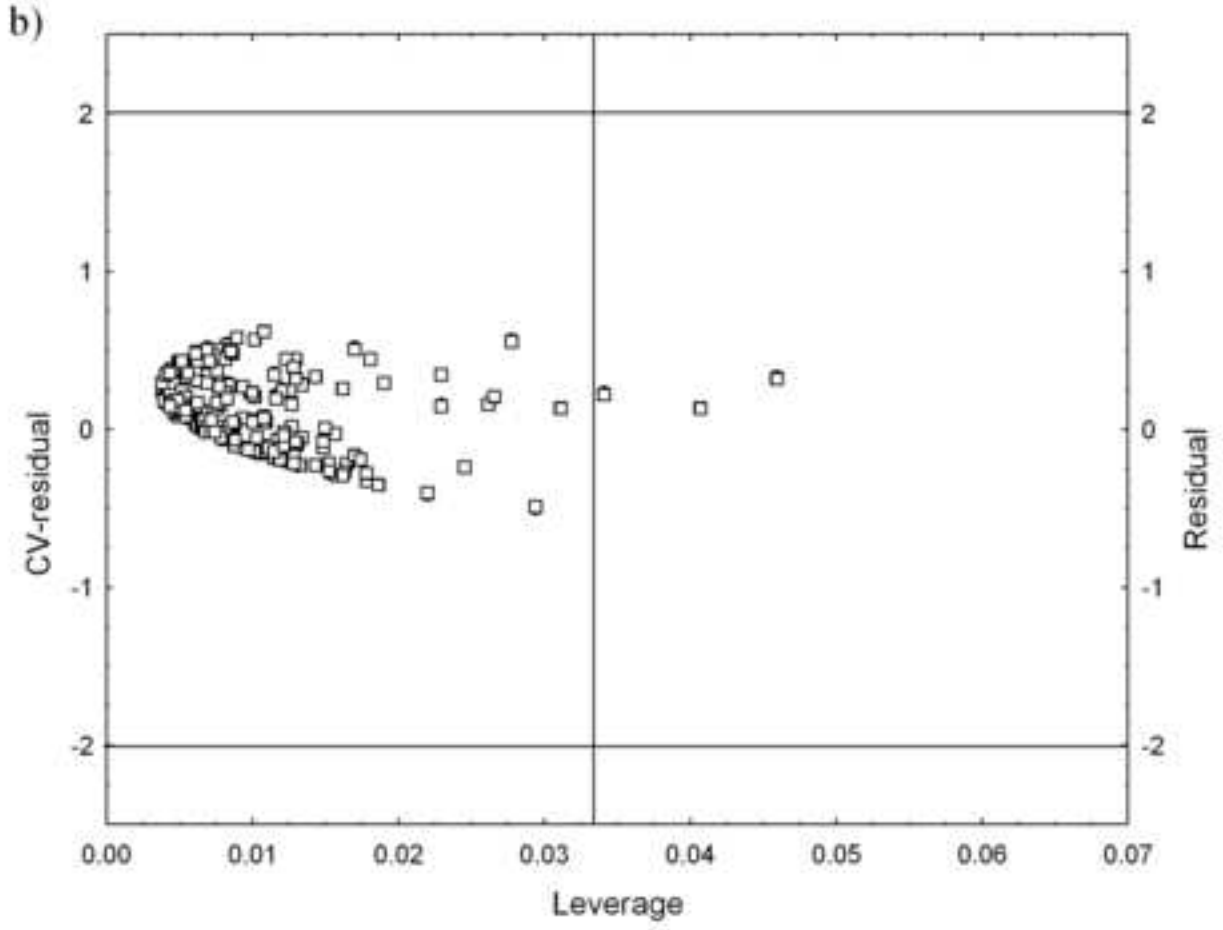
ACCEPTED MANUSCRIPT



Figure

ACCEPTED MANUSCRIPT

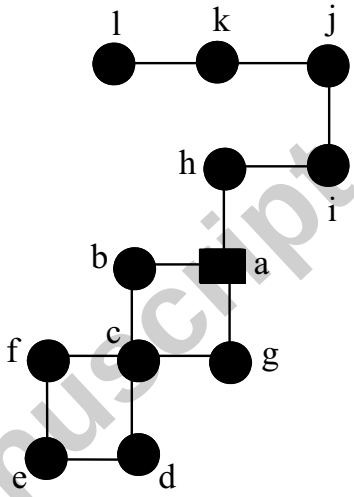




ACCEPTED MANUSCRIPT

Table 1. Description of Nodes, Aminoacid Sequence, Coordinates and Stochastic Matrix for a HP Lattice Network.

n	Aminoacids	X	Y	HP Lattice Network
a	Ser1	0	0	
	Cys9	0	0	
b	Hys2	-1	0	
c	Leu3	-1	-1	
	Glu7	-1	-1	
d	Val4	-1	-2	
e	Hys5	-2	-2	
f	Asn6	-2	-1	
g	Asp8	0	-1	
h	Tyr10	0	1	
i	Glu11	1	1	
j	Tyr12	1	2	
k	Hys13	0	2	
l	Lys14	-1	2	



$$\begin{bmatrix} {}^{1}p_{aa} & {}^{1}p_{ab} & 0 & 0 & \ldots\ldots\ldots\ldots 0 & 0 \\ {}^{1}p_{ba} & {}^{1}p_{bb} & {}^{1}p_{bc} & 0 & \ldots\ldots\ldots\ldots 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots\ldots\ldots\ldots & {}^{1}p_{kk} & {}^{1}p_{kl} \\ 0 & 0 & \ldots\ldots\ldots\ldots & {}^{1}p_{lk} & {}^{1}p_{ll} \end{bmatrix}$$

Accepted manuscript

ACCEPTED MANUSCRIPT

Table 2. Proteins Introduced in the Model and their *a Posteriori* Probabilities (P). 62 of 69 Positive Cases are well Evaluated (89.86% of Good Classification). 166 of 200 Negative Cases are well Evaluated (83.00% of Good Classification). P (jackknife) is the *a Posteriori* Probability Extracted from the Jackknife Test.

Nº	Protein	P	P(jackknife)	Nº	Protein	P	P(jackknife)
HCC related PROTEINS (we give the protein gene name)							
1	ABCA1	0.57	0.56	36	MCP	0.71	0.70
2	ACSL5	0.99	0.99	37	MGC33407	0.95	0.95
3	ADAM29	0.91	0.91	38	MKRN3	0.64	0.63
4	ADAMTS15	0.91	0.91	39	MLL3	0.91	0.90
5	ADAMTS18	0.93	0.93	40	MMP2	0.70	0.70
6	ADAMTSL3	0.79	0.78	41	NF1	0.96	0.96
7	APC	0.65	0.65	42	OBSCN	0.95	0.95
8	C6orf29	0.93	0.92	43	P2RX7	0.84	0.84
9	C10orf137	0.27	0.26	44	P2RY14	0.96	0.96
10	C15orf2	0.94	0.94	45	PHIP	0.83	0.82
11	CD109	0.91	0.91	46	PKHD1	0.87	0.86
12	CD248	0.40	0.39	47	PKNOX1	0.94	0.94
13	CHL1	0.78	0.78	48	PRKD1	0.98	0.98
14	CNTN4	0.61	0.60	49	PTPRD	0.81	0.81
15	CSMD3	0.80	0.79	50	PTPRU	0.93	0.93
16	EPHA3	0.62	0.62	51	RET	0.79	0.79
17	EPHB6	0.89	0.89	52	RUNX1T1	0.66	0.66
18	ERCC6	0.24	0.23	53	SCN3B	0.84	0.84
19	EVL	0.96	0.96	54	SDBCAG84	0.84	0.84

Accepted manuscript

ACCEPTED MANUSCRIPT

20	EYA4	0.99	0.99	55	SEC8L1	0.91	0.91
21	FBXW7	0.94	0.94	56	SFRS6	0.44	0.43
22	GALNS	0.90	0.90	57	SLC29A1	0.98	0.98
23	GNAS	0.43	0.41	58	SMAD2	0.72	0.71
24	GUCY1A2	0.77	0.77	59	SMAD3	0.86	0.86
25	HAPLN1	0.93	0.93	60	SMAD4	0.73	0.73
26	HIST1H1B	0.99	0.99	61	SYNE1	0.96	0.96
27	K6IRS3	0.88	0.88	62	TBX22	0.92	0.92
28	KCNQ5	1.00	0.99	63	TCF7L2	0.73	0.73
29	KIAA1409	0.81	0.80	64	TGFBR2	0.89	0.89
30	KRAS	0.03	0.02	65	TP53	0.97	0.97
31	LGR6	0.99	0.99	66	TTLL3	0.97	0.97
32	LMO7	0.73	0.73	67	UHRF2	0.95	0.95
33	LOC157697	0.97	0.97	68	UQCRC2	0.89	0.89
34	LRP2	0.84	0.83	69	ZNF442	0.97	0.97
35	MAP2	0.30	0.27				

No-HCC related PROTEINS (we give the PDB ID)

70	1A0K	0.02	0.02	170	1PBK	0.03	0.03
71	1A0S P	0.52	0.53	171	1PHR	0.01	0.01
72	1A0S Q	0.52	0.53	172	1PMT	0.64	0.64
73	1A0S R	0.52	0.53	173	1PUD	0.52	0.53
74	1A1X	0.01	0.01	174	1QF7 B	0.03	0.03
75	1A2B	0.23	0.23	175	1QGH A	0.03	0.03
76	1A2J	0.08	0.08	176	1QMG D	0.12	0.12

Accepted manuscript

ACCEPTED MANUSCRIPT

77	1A3Z	0.24	0.24	177	1QTO A	0.02	0.02
78	1A8P	0.15	0.16	178	1RCB	0.02	0.02
79	1A44	0.02	0.02	179	1RLW	0.03	0.03
80	1A45	0.05	0.05	180	1SKZ	0.01	0.01
81	1A62	0.01	0.01	181	1SVP A	0.08	0.08
82	1AAZ A	0.00	0.00	182	1TYB E	0.72	0.73
83	1AAZ B	0.00	0.00	183	1UBV	0.60	0.60
84	1ALU	0.31	0.32	184	1UP1	0.06	0.06
85	1AQE	0.00	0.00	185	1VNC	0.67	0.68
86	1AVU	0.16	0.16	186	1WHO	0.00	0.00
87	1B0L A	0.78	0.78	187	1WU3 I	0.14	0.15
88	1B9W A	0.00	0.00	188	1WWB X	0.00	0.00
89	1BAS	0.04	0.04	189	1XAN	0.58	0.60
90	1BD8	0.12	0.12	190	1XIB	0.17	0.18
91	1BKB	0.12	0.13	191	1XJO	0.34	0.34
92	1BV1	0.04	0.04	192	1XSO A	0.00	0.00
93	1BXM	0.01	0.01	193	1XSO B	0.00	0.00
94	1C1L A	0.05	0.05	194	1YGH A	0.10	0.10
95	1C5E A	0.01	0.01	195	1YHB	0.00	0.00
96	1CC7 A	0.00	0.00	196	1YTT A	0.02	0.02
97	1CEN	0.51	0.52	197	1ZIN	0.52	0.52
98	1CL0 A	0.43	0.43	198	1ZRM	0.06	0.06
99	1COT	0.00	0.00	199	2A0B	0.03	0.03
100	1CPM	0.52	0.53	200	2AXE	0.84	0.85
101	1CXA	0.01	0.01	201	2BNH	0.07	0.07

Accepted manuscript

ACCEPTED MANUSCRIPT

102	1DIX A	0.71	0.72	202	2CI2 I	0.00	0.00
103	1DK8 A	0.04	0.04	203	2DHN	0.01	0.01
104	1DOT	0.63	0.64	204	2DUB F	0.29	0.32
105	1DT1 A	0.02	0.02	205	2EQL	0.03	0.03
106	1DUC	0.03	0.03	206	2ERK	0.46	0.47
107	1DY3 A	0.08	0.09	207	2FAL	0.06	0.06
108	1E2U A	0.58	0.60	208	2FD2	0.00	0.00
109	1E7L B	0.02	0.02	209	2FHA	0.08	0.08
110	1EAJ B	0.01	0.01	210	2FIT	0.09	0.09
111	1ED1 A	0.03	0.03	211	2FKE	0.00	0.00
112	1EJB E	0.05	0.05	212	2FUA	0.14	0.15
113	1EL5 A	0.23	0.23	213	2GAC A	0.01	0.01
114	1ET6 B	0.25	0.25	214	2GAC B	0.01	0.01
115	1F0M A	0.00	0.00	215	2GAC C	0.01	0.01
116	1F47 B	0.06	0.06	216	2GAC D	0.01	0.01
117	1F83 A	0.23	0.24	217	2GDM	0.17	0.18
118	1FHG A	0.04	0.04	218	2GLT	0.26	0.26
119	1FYH D	0.35	0.36	219	2HPR	0.00	0.00
120	1G0W A	0.31	0.32	220	2IMM	0.00	0.00
121	1G6N B	0.07	0.07	221	2IMN	0.00	0.00
122	1G7C B	0.01	0.01	222	2INT	0.02	0.02
123	1GUP A	0.24	0.24	223	2IZA	0.01	0.01
124	1H4Y B	0.01	0.01	224	2LIV	0.65	0.66
125	1HI3 A	0.05	0.05	225	2MAD L	0.11	0.11
126	1HMY	0.24	0.25	226	2MBR	0.21	0.22

Accepted manuscript

ACCEPTED MANUSCRIPT

127	1HQ3 D	0.01	0.01	227	2PF1	0.13	0.13
128	1HUS	0.12	0.12	228	2RSL A	0.08	0.08
129	1HX3 A	0.16	0.17	229	2TCT	0.24	0.24
130	1HX3 B	0.16	0.17	230	2TDT	0.34	0.35
131	1I81 E	0.00	0.00	231	2TGI	0.00	0.00
132	1ILR 1	0.05	0.05	232	2TIR	0.02	0.02
133	1INO	0.44	0.44	233	2TMY	0.02	0.02
134	1IOP	0.12	0.12	234	2TN4	0.06	0.06
135	1IRD A	0.26	0.26	235	2TS1	0.83	0.83
136	1IRD B	0.02	0.02	236	2UBP A	0.00	0.00
137	1IUZ	0.01	0.01	237	2UBP B	0.01	0.01
138	1IXG	0.56	0.56	238	2UBP C	0.48	0.48
139	1JAH	0.02	0.02	239	2UKD	0.73	0.75
140	1JDO	0.11	0.11	240	2WBC	0.81	0.83
141	1JEH A	0.08	0.08	241	2WRP R	0.00	0.00
142	1JEH B	0.08	0.08	242	3BC2	0.10	0.10
143	1JLM	0.21	0.21	243	3BIR	0.02	0.02
144	1JMW A	0.06	0.06	244	3BLM	0.18	0.18
145	1JWO A	0.00	0.00	245	3CLN	0.07	0.07
146	1K89	0.06	0.07	246	3CSC	0.69	0.69
147	1KAA	0.02	0.02	247	3CYR	0.00	0.00
148	1KDJ	0.01	0.01	248	3ENG	0.70	0.70
149	1KLO	0.57	0.59	249	3GBP	0.55	0.56
150	1KMB 1	0.08	0.08	250	3KAR	0.56	0.56
151	1KVA	0.10	0.10	251	3PNP	0.37	0.38

Accepted manuscript

ACCEPTED MANUSCRIPT

152	1KVW	0.03	0.03	252	3PYP	0.01	0.01
153	1LAM	0.67	0.68	253	3RHN	0.04	0.04
154	1LE4	0.15	0.15	254	3RUB S	0.00	0.01
155	1LIT	0.29	0.30	255	3SEB	0.11	0.11
156	1LOP A	0.13	0.13	256	3SSI	0.01	0.01
157	1LOU A	0.01	0.01	257	3VUB	0.00	0.00
158	1MAR	0.35	0.36	258	4AIG	0.56	0.57
159	1MHO	0.00	0.00	259	4LVE A	0.01	0.01
160	1MOL A	0.00	0.00	260	4LZM	0.06	0.06
161	1MOL B	0.00	0.00	261	4PAH	0.37	0.37
162	1MRG	0.66	0.67	262	5EAU	0.73	0.74
163	1MUP	0.01	0.01	263	6TAA	0.88	0.89
164	1MZM	0.00	0.00	264	7ACN	0.68	0.68
165	1NCX	0.13	0.13	265	7ATJ A	0.69	0.69
166	1NFO	0.43	0.43	266	7PAZ	0.03	0.03
167	1NNA	0.41	0.42	267	7PCY	0.01	0.01
168	1OHJ	0.38	0.38	268	8CHO	0.02	0.02
169	1OPC	0.01	0.01	269	451C	0.00	0.00

Accepted manuscript

ACCEPTED MANUSCRIPT

Table 3. Description of the Models Varying the Cross-validation Series.

	CV1	CV2	CV3	CV4
N	202	202	202	202
Wilks' λ	0.55	0.51	0.56	0.53
F(Fisher)	80.61	95.38	76.70	87.11
p	<0.00	<0.00	<0.00	<0.00
%active training	88.24	90.38	88.00	90.20
%active cross-validation	94.44	88.24	94.74	88.89
%inactive training	84.11	82.00	81.58	84.11
%inactive cross-validation	79.59	86.00	87.50	81.63
%Total	84.76	84.76	84.76	85.13

Accepted manuscript