Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices

Cristian Robert Munteanu, Alexandre L. Magalhães, Eugenio Uriarte, Humberto González-Díaz

www.elsevier.com/locate/yjtbi

Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices

Cristian Robert Munteanu, Alexandre L. Magalhães, Eugenio Uriarte, Humberto González-Díaz

PII: S0022-5193(08)00614-0
DOI: doi:10.1016/j.jtbi.2008.11.017
Reference: YJTBI5379

To appear in: *Journal of Theoretical Biology*

Received date: 22 July 2008
Revised date: 11 November 2008
Accepted date: 22 November 2008

Cite this article as: Cristian Robert Munteanu, Alexandre L. Magalhães, Eugenio Uriarte and Humberto González-Díaz, Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices, *Journal of Theoretical Biology* (2008), doi:10.1016/j.jtbi.2008.11.017

Multi-Target QPDR Classification Model for Human Breast and Colon Cancer-Related Proteins using Star Graph Topological Indices

CRISTIAN ROBERT MUNTEANU,[1] ALEXANDRE L. MAGALHÃES,[1] EUGENIO URIARTE,[2] HUMBERTO GONZÁLEZ-DÍAZ,[2,*]

[1]*REQUIMTE/Faculty of Science, Chemistry Department, University of Porto 4169-007, Portugal,* muntisa@gmail.com, almagalh@fc.up.pt

[2] *Unit of Bioinformatics & Connectivity Analysis (UBICA), Institute of industrial Pharmacy, and Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15782, Spain,* humberto.gonzales@usc.es , *eugenio.uriarte@usc.es*

Abstract. The cancer diagnostic is a complex process and, sometimes, the specific markers can interfere or produce negative results. Thus, new simple and fast theoretical models are required. One option is the complex network graphs theory that permits us to describe any real system, from the small molecules to the complex genetic, neural or social networks by transforming real properties in topological indices. This work converts the protein primary structure data in specific Randic's star networks topological indices using the new Sequence to Star Networks (*S2SNet*) application. A set of 1054 proteins were selected from previous works and contains proteins related or not with two types of cancer, HBC (human breast cancer) and HCC (human colon cancer). The General Discriminant Analysis method generates an input-coded multi-target classification model with the training/predicting set accuracies of 90.0% for the Forward Stepwise model type. In addition, a protein subset was modified by single amino acid mutations with higher log-odds PAM250 values and tested with the new classification if can be related with HBC or HCC. In conclusion, we shown that, using simple input data such is the primary protein sequence and the simples linear analysis, it is possible to obtain accurate classification models that can predict if a new protein related with two types of cancer. These results promote the use of the S2SNet in clinical proteomics.

Keywords: input-coded multi-target QPDR, star graph, cancer theoretical model, clinical proteomics, GDA method.

---

*Corresponding author. Unit of Bioinformatics & Connectivity Analysis (UBICA), and Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela 15782, Spain. Email: humbertogd@gmail.com, Tel: +34-981563100, Fax: +34-981594912.

1. Introduction

Cancer is a leading cause of death worldwide, accounted for around 13% of all deaths in 2007 (WHO, 2008). Two of the leading types of cancer are the human breast cancer (HBC) and the human colon cancer (HCC). The estimated new cancer cases and deaths in United States for 2008 shows that HBC will affect 26% of the women (15% will die) and HCC will involve 10% of the men/women (8% men and 9% women will die) (Jemal et al., 2008). Therefore, simple and fast theoretical method can be very useful in the detection of cancer diseases.

The actual work will use the protein Quantitative Proteome-Disease Relationship (QPDR) (Ferino et al. 2008), similar to Quantitative structure-activity relationship (QSAR) (Devillers and Balaban, 1999). QPDR is one of the widely used analyse for predicting the protein properties and, in the present study, is using the macromolecular descriptors, named topological indices (TIs), obtained with the graph theory. The branch of mathematical chemistry dedicated to encode the DNA/protein information in graph representations by the use of the TIs has become an intense research area (Agüero-Chapin et al., 2006; Bielinska-Waz et al., 2007; Liao and Wang, 2004; Liao and Ding, 2005; Randic, 2000; Randic and Basak, 2001; Randic and Balaban, 2003; Randic et al., 2000). The graphic approaches of the biological systems study can provide useful insights in QSAR studies (González-Díaz et al., 2006; González-Díaz et al., 2007c; Prado-Prado et al., 2008), protein folding kinetics (Chou, 1990), enzyme-catalyzed reactions (Chou, 1989; Chou and Forsen, 1980; Chou and Liu, 1981; Kuzmic et al., 1992), inhibition kinetics of processive nucleic acid polymerases and nucleases (Althaus et al., 1993a; Althaus et al., 1993b; Althaus et al., 1994; Althaus et al., 1996; Chou et al., 1994), DNA sequence analysis (Qi et al., 2007), anti-sense strands base frequencies (Chou et al., 1996), analysis of codon usage (Chou and Zhang, 1992; Zhang and Chou, 1994) and in complicated network systems investigations (Diao et al., 2007; Gonzalez-Diaz et al., 2008; González-Díaz et al., 2007a). Recently, the "cellular automaton image" (Wolfram, 1984; Wolfram, 2002) has also been applied to study hepatitis B viral infections (Xiao et al., 2006a), HBV virus gene missense mutation (Xiao et al., 2005b), and visual analysis of SARS-CoV (Gao et al., 2006; Wang et al., 2005), as well as representing complicated biological sequences (Xiao et al., 2005a) and helping to identify protein attributes (Xiao and Chou, 2007; Xiao et al., 2006b). We have chosen the TIs for these QPDR models based on the previous work results with similar QSAR/QPDR models. Even if the TIs can not be always interpreted, they

demonstrate to encode the information that permits to create accurate QSAR/QPDR models.

Other interesting fields to apply the graph theory are the oncology and clinical proteomics. A classification model for discriminating prostate cancer patients from control group with connectivity indices where constructed by Gonzales et al. (González-Díaz et al., 2007b). Vilar's group designed a QSAR model for alignment-free prediction of HBC biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks (Vilar et al., 2008).

The actual work is proposing a new cancer / non-cancer classification model based on protein embedded / non-embedded Star Graph TIs such are the trace of connectivity matrices, Harary number, Wiener index, Gutman index, Schultz index, Moreau-Broto indices, Balaban distance connectivity index, Kier-Hall connectivity indices and Randic connectivity index. This classification can predict two types of cancer: HBC and HCC. The primary protein sequence is transformed in connectivity Star Graph's TIs that are used by a statistical linear method in order to construct an input-coded multi-target classification model.

2. Materials and Methods

2.1. Protein Set

Two sets of protein primary sequences are used: a set of 189 HBC/HCC cancer proteins (Sjoblom et al., 2006) and 865 non-cancer proteins (Dobson and Doig, 2005; Dobson et al., 2004). The list of cancer-related proteins in our work is the same with the list obtained by the Sjoblom group after the experimental analysis of 13,023 genes in 11 breast and 11 colorectal cancers.

2.2. Star Graph Topological Indices

Each protein sequence was transformed in a star graph, where the amino acids are the vertices (nodes), connected in a specific sequence by the peptide bonds. The star graph is a special case of trees with N vertices where one has got N-1 degrees of freedom and the remaining N-1 vertices have got one single degree of freedom (Harary, 1969). Each of the 20 possible branches ("rays") of the star contains the same amino acid type and the star centre is a non-amino acid vertex.

A protein can be represented by diverse forms of graphs, which can be associated with

distinct distance matrices. The best method to construct a standard star graph is the following: each amino acid/vertex holds the position in the original sequence and the branches are labelled by alphabetical order of the 3-letter amino acid code (Randic et al., 2007). The graph is embedded if the initial sequence connectivity in the protein chain is included. **Figures 1A** and **1B** present the non-embedded/embedded star graphs of PRPS1 using the alphabetical order of one-letter amino acid code. Thus, the primary structure of protein chains are transformed in the correspondent Star graphs invariant TIs. The resulted graphs are not depending on the 3-dimensional structure or the shape of the protein.

Figure 1A comes about here

Figure 1B comes about here

The comparison of the graphs is made by using the corresponding connectivity matrix, distance matrix and degree matrix. The matrices of the connectivity in the sequence and in the star graph are combined in the case of the embedded graph. These matrices and the normalized ones are the base of the TIs calculation.

The conversion of the amino acid sequences in star graph TIs was made by using Sequence to Star Networks (S2SNet) application, developed by our group (Munteanu and Gonzáles-Diáz, 2008). S2SNet is based on wxPython (Rappin and Dunn, 2006) for the GUI application and has *Graphviz* (Koutsofios and North, 1993) as a graphics back-end. The present calculations are characterized by embedded and non-embedded TIs, no weights, Markov normalization and power of matrices/indices (n) up to 5. The results file contains the following TIs (Todeschini and Consonni, 2002):

- Trace of the n connectivity matrices (Tr_n) or the spectral moments:

$$Tr_n = \sum_i (M^n)_{ii}, \quad (1)$$

where n = 0 – power limit, M = graph connectivity matrix ($i*i$ dimension); ii = i^{th} diagonal element;

- Harary number (H) or the reciprocal distance sum index:

$$H = \sum_{i<j} m_{ij}/d_{ij}, \quad (2)$$

where d_{ij} are the elements of the distance matrix and m_{ij} are the elements of the M connectivity matrix;

- Wiener index (W) or the sum of the numbers of edges in the shortest paths in a graph between all pairs of amino acids in a protein:

$$W = \sum_{i<j} d_{ij}, \quad (3)$$

- Gutman topological index (S_6):

$S_6 = \sum_{ij} deg_i * deg_j / d_{ij}$, (4)

where deg_i are the elements of the degree matrix;

- Schultz topological index (non-trivial part) (S):

$S = \sum_{i<j} (deg_i + deg_j) * d_{ij}$, (5)

- Balaban distance connectivity index (J) or average distance sum connectivity index (measures the graph ramification):

$J = (edges - nodes + 2) * \sum_{i<j} m_{ij} * \text{sqrt}(\sum_k d_{ik} * \sum_k d_{kj})$,

(6)

where $nodes$+1 = AA numbers/node number in the Star Graph + origin, $\sum_k d_{ik}$ is the node distance degree;

- Kier-Hall connectivity indices (^nX):

$^0X = \sum_i 1 / \text{sqrt}(deg_i)$, (7)

$^2X = \sum_{i<j<k} m_{ij} * m_{jk} / \text{sqrt}(deg_i * deg_j * deg_k)$, (8)

$^3X = \sum_{i<j<k<m} m_{ij} * m_{jk} * m_{km} / \text{sqrt}(deg_i * deg_j * deg_k * deg_m)$, (9)

$^4X = \sum_{i<j<k<m<o} m_{ij} * m_{jk} * m_{km} * m_{mo} / \text{sqrt}(deg_i * deg_j * deg_k * deg_m * deg_o)$, (10)

$^5X = \sum_{i<j<k<m<o<q} m_{ij} * m_{jk} * m_{km} * m_{mo} * m_{oq} / \text{sqrt}(deg_i * deg_j * deg_k * deg_m * deg_o * deg_q)$, (11)

- Randic connectivity index (^1X):

$^1X = \sum_{ij} m_{ij} / \text{sqrt}(deg_i * deg_j)$, (12)

These TIs and other derivate ones will be used in the next step to construct a cancer / non-cancer classification model by linear statistical methods.

2.3. Statistical Analysis

An input-coded multi-target classification model was created with General Discriminant Analysis (GDA) method (Kowalski and Wold, 1982; Van Waterbeemd, 1995), STATISTICA 6.0 package (StatSoft.Inc., 2002). This model can predict if a protein is HBC or HCC-related using a single equation. For this reason, in addition to the 30 star graph embedded and non-embedded TIs are introduced other two types of continuous predictors (attributes) encoded specific information about each cancer types as following: 30 products of the HBC/HCC cancer probability with the embedded / non-embedded TIs (**p**TI = $\text{prob}_{\text{HBC/HCC}}$*TI) and 30 differences between the same TIs and the average of the TIs for each type of cancer [**d**TI=TI – $\text{average(TI)}_{\text{HBC/HCC}}$]. The cancer probabilities represent the fractions of proteins HBC/HCC-related from the entire Sjöblom's proteins (cancer proteins) and have values of 0.639 (HBC) and 0.361 (HCC). For each protein there are two cases corresponding to both types of cancer. The

dependent variable (CancerOrNot) takes 1 for cancer and 0 for non-cancer and the cross-validation variable (*CV*) has two values (*train* and *val*). The best cross-validation methods to examine a predictor are the following: independent dataset test, subsampling test, and jackknife test (Chou and Zhang, 1995). Chou and Shen have shown that only the jackknife test has the least arbitrariness (Chou and Shen, 2007; Chou and Shen, 2008). Thus, the jackknife test has been increasingly used by investigators to examine the accuracy of various predictors (Chen and Li, 2007a; Chen and Li, 2007b; Diao et al., 2007; Ding et al., 2007; Jiang et al., 2008; Li and Li, 2008; Lin, 2008; Niu et al., 2006; Xiao and Chou, 2007; Zhang et al., 2008; Zhou et al., 2007). In the actual work, the independent data test is used by splitting the data at random in a training series (*train*, 75%) used for model construction and a prediction one (*val*, 25%) for model validation (the CV column is filled by repeating 6 *train* and 2 *val*). All independent variables are standardized prior to model construction.

The general QPDR formula contains embedded and non-embedded TIs, **p**TIs and **d**TIs:

$$C/nC\text{-score} = c_0 + \sum_{i=1->n} c_i * TI_i + \sum_{j=n->m} c_j * \mathbf{p}TI_j + \sum_{k=m->o} c_k * \mathbf{d}TI_k, \qquad (13)$$

where *C/nC*-score is the continue score value for the cancer / non-cancer classification (HBC or HCC), c_1-c_n are the TIs coefficients (n=number of TIs), c_n-c_m **p**TIs coefficients (n<m; m-n= number of **p**TIs), c_m-c_o **d**TIs coefficients (m<o; o-m= number of **d**TIs) and c_0 is the independent term. We inspected the percentage of good classification and the number of variables to be explored in order to avoid over-fitting or chance correlation. The *Forward* model type was tested for the embedded, non-embedded and both data, including TIs, **p**TIs, **d**TIs and all indices.

In addition, the Dobson's set is use to select a subset of 61 non-cancer proteins with cancer probability between 0.3 and 0.5 in order to proceed 17 single amino acid mutations with log-odds PAM250 (Dayhoff, 1978) greater or equal with 2 (see **Table 1**). The best classification model predicted the probability of presence in HBC/HCC cancer for any of these mutated proteins and the results were analysed with two-way joining clustering analysis method (tw-JCA) from STATISTICA (StatSoft.Inc., 2002).

Table 1 comes about here

3. Results and Discussions

Fifteen classification models were tested with the aim of finding the best GDA equation which is able to discriminate between proteins related with HBC and HCC. The

attributes include 30 embedded/non-embedded Star Graph TIs obtained with S2SNet application and other 60 composed predictors, **p**TIs and **d**TIs. The values obtained for the training/predicting accuracies with the *Forward Stepwise* method are presented in **Table 2**.

Table 2 comes about here

The *Forward Stepwise* selection variable method conjugated with the embedded TIs & **d**TIs provides the best results for our data set with values of correctly classified compounds of 89.9%, 90.3% and 90.0% for the training, cross-validation and full sets, respectively, and using only six/five parameters/variables (**Eq. 14**). The embedded TIs have the name of the non-embedded ones plus "**e**" as suffix. The simple linear mathematical form of the model has been chosen in the absence of prior information.

C/nC-score = -4.4+1.7**Tr3***e**+124.8**S***e**-126.5***d***J***e**+48.6***d***X2***e**-45.9**X5***e**, (14)

N=2102, R_c=0.54, *U*=0.70, *F*=132.20, p<0.001,

where *N* is the number of cases (*C&nC*), R_c is the canonical regression coefficient, *U* is the Wilk's statistics, *F* is the Fisher's statistics and p is the p-level (probability of error).

The above results are typically considered as excellent in the literature for LDA-QPDR/QSAR models (Castillo-Garit et al., 2008; Estrada and Molina, 2001; Marrero-Ponce et al., 2004; Morales et al., 2006; Vilar et al., 2008). In order to check the variation of this model with the training/cross-validation sets, we carried on a cross-validation study by using ten totally random sets, including the initial one from the actual model (with the same 75% training and 25% cross-validation). The classification values are presented in **Table S1** from the supplementary material and show an average of 90.2% for training and 89.2% for cross-validation. These values demonstrate the stability of the model with the selection of the classification sets.

In order to illustrate the performance of the approach when applied to a single set of cancer related proteins (e.g. either breast or colon), we obtained two equations, one for HBC and other for HCC. Therefore, we have to consider that the **Eq.14** represents an input-coded multi-target classification model that can evaluate if a protein is HBC or HCC-related by using the HBC or HCC average *J***e** and *X2***e** values (contained in the **d***J***e** and **d***X2***e** differences). **Eq. 14** can be reduced to two different equations, one for each type of cancer (HBC and HCC):

HBC/nHBC-score = -19.8+1.7**Tr3***e**+124.8**S***e**-*J***e**+0.2**X2***e**-45.9**X5***e**, (14a)

HCC/nHCC-score = -20.8+1.7**Tr3***e**+124.8**S***e**-*J***e**+0.2**X2***e**-45.9**X5***e**, (14b)

The detailed classification results for each type of cancer obtained with **Eq. 14a** and

Eq.14b are presented in **Table 3**.

Table 3 comes about here

A similar input-coded multi-target classification model was obtained by using the *Forward Stepwise* method and the embedded **p**TIs and provides values of correctly classified compounds of 90.3%, 91.0% and 90.5% for the training, cross-validation and full sets, respectively (using seven/six parameters/variables) (**Eq. 15**).

C/nC-score = -4.1-118.6***p***Tr0***e*+80.7***p***Tr2***e*+1.4***p***Tr3***e*+100.3***p***S***e*
-101.4***p***J***e*+39.7***p***X2***e*, (15)

N=2102, R_c=0.58, *U*=0.66, *F*=135.08, p<0.001,

In order to evaluate if a protein is HBC or HCC-related, it is necessary to use the HBC or HCC probability inside the **p**TIs products. The classification values obtained for the individual equations are presented in **Table 3**. The equations obtained are the following:

HBC/nHBC-score = -5.6-0.3**Tr0***e*+0.8**Tr2***e*+0.6**Tr3***e*+0.2**X2***e*, (15a)

HCC/nHCC-score = -5.6-0.2**Tr0***e*+0.5**Tr2***e*+0.3**Tr3***e*+0.1**X2***e*, (15b)

Eq. 14 and **Eq.15** show similar results when the input data is containing probability of cancer (products with TIs) or the TIs averages for each type of cancer (differences with TIs). In general, in the case of embedded, non-embedded and both indices, we obtained better results with **d**TIs compared with the **p**TIs (not mixed with the original TIs). This difference can be explained by a superior recover of the cancer-related protein sequence information in the case of the differences between the original TIs and the average of them for each type of cancer (**d**TIs) compared with the products of the original TIs and the cancer type probability (**p**TIs). Thus, we can conclude that the average of star graph structure for each type of cancer (**d**TIs) is described better the actual QPDR model compared with the composition of the data sets for each type of cancer that generates the cancer probabilities. In addition, Table 2 shows that better results are obtained using the original TIs and the derived ones (**p**TIs and **d**TIs) compared with the isolated TIs/**p**TIs/**d**TIs. This difference can be explained be the fact that each set of indices can contains different parts of the protein information that is cancer-related. Therefore, the use of all these indices will sum all this information in a better QPDR model.

Another interesting aspect is the type of the indices (original or derived from the original) that are more frequent in all models presented in **Table S2** from supplementary material. Thus, we can observe the importance of the Wiener index (W) and Kier Hall connectivity index X5 for the models based on the non-embedded TIs. The embedded TIs models contain more frequent the trace of the graph/sequence connectivity matrixes

Tr3 and the non-trivial part of the Schultz topological index S (W is based on the distance matrix, X5 and S on the degree matrix, and Tr3 on the connectivity matrix). The most important type of index that is present in both embedded and non-embedded TI equations is J, the Balaban distance connectivity index based on the node distance degree information. In order to compare two equations with the same number of TIs, we have chosen the embedded models with **p**TI**e** and {TI**e**, **p**TI**e**} that contain 6 variables and reduced the common terms (based on Tr3, S and J). Thus, we can observe that the addition of the TI**e** to the **p**TI**e** will shift the preference from the low order traces (**p**Tr0**e**, **p**Tr2**e**) and Kier-Hall index (**p**X2**e**) to high order trace (Tr5**e**), Harary number (H**e**) and Gutman topological index (S6**e**).

The first embedded TIs & **d**TIs model was chosen to estimate the cancer probability for proteins mutants of non-cancer-related proteins. These values were analyzed with tw-JCA using 61 mutated proteins and 17 types of single amino acid mutations. In the case of HBC, we obtained 215 data groups, called input blocks. To detect the larger variability regions (mutants) we computed a tw-JCA partition of input blocks (rearrange of blocks) setting the threshold value of variability at StDv/2 (see **Figure 2**). The value obtained was 0.059. The 215 input blocks are regrouped, for similarity, into 11 output blocks (see **Tables S3** and **S4** in the supplementary material). We can observe that the proteins with number 24 to number 48 are very susceptible to become HBC-related proteins for all studied mutations. The plot corresponding to the reduced values of the reordered data matrix (**Table S4**) is presented in **Figure 3**. On the other hand, we carried out the same study for the HCC mutated proteins and found different susceptible proteins, with visible lower probability to be HCC-related (**Figure 4**). The 184 input blocks were regrouped, for similarity, into 11 output blocks (StDv/2=0.050) (see **Tables S5** and **S6** in the supplementary material). The reduced data from **Table S6** are presented as a plot in **Figure 5**. The tw-JCA partition obtained in this way is statistically significant as reported by other authors that used this method to reach similar goals (Ferino et al., 2008).

Figure 2 comes about here

Figure 3 comes about here

One interesting non-cancer chain protein is 1QRK B, the human coagulation factor XIII with strontium bound in the ion site (Fox et al., 1999), with eight single amino acid mutations that present HBC probability up to 71% as following: 70.8% for V->L, 68.8% for V->I, 62.0% for L->I, 59.3% for D->N, 58.3% for E->Q, 55.9% for F->L,

54.8% for E->D and 51.0% for V->M. The most persistent mutation (log-odd PAM250=4), valine (V) to isoleucine (M), can be considered as the most dangerous one. The main calcium/strontium binding site within each monomer involves the main chain oxygen atom of Ala-457, and also the side chains from residues Asn-436, Asp-438, Glu-485, and Glu-490. The mutations of Glu (E) in Q and D can affect the capacity of binding metals and the normal biological activity. This coagulation factor XIII is a transglutaminase which stabilises blood clots by covalently cross-linking fibrin, being essential for normal haemostasis. FXIII deficiency due to the genetic mutations results in a life-long bleeding disorder with added complications in wound healing and tissue repair (Anwar et al., 1998). In addition, the abundant fibrinogen present in the tumor connective tissue might contribute to the structural integrity of breast or colon tumor tissues (Costantini et al., 1991; Takahashi et al., 2000; Yee et al., 1994). We can observe that, in general, the natural mutations with higher PAM250 values are less frequent even for 1QRK B (Y->F with PAM250 of 7 is absent) because we can not create a direct relation between the PAM250 natural amino acid mutation frequency and the influence of the mutations in these types of cancer.

Figure 4 comes near here

Figure 5 comes near here

The probability for a cancer-related protein to turn into a non-cancer one was studied too. For each type of cancer, ten HBC/HCC-related proteins where mutated using the same PAM250 values. The tw-JCA plots are presented in **Figure 6** (for HBC) and **Figure 7** (for HCC), and correspond to data in **Table S7** and **Table S8** from the supplementary material. The results did not show important probability to obtain a HBC/HCC-related protein by using single PAM250 natural mutations. Activin beta E (INHBE, C_5) has the highest probability (around 50%) to turn into a HBC-related protein after almost all the mutations (**Figure 6** and **Table S7**).

Figure 6 comes near here

Figure 7 comes near here

4. Conclusions

This study is proposing two cancer / non-cancer input-coded multi-target classification models for HBC and HCC using the Star Network TIs of the protein amino acid sequences. The results prove the excellent predictive ability (90.0%) of the simple and fast Star Network TIs and GDA statistics linear models in the case of the actual protein

model. In addition, the prediction of cancer probability for mutated proteins was calculated. The human coagulation factor XIII (1QRK B), that normally do not generate HBC, if suffer several mutations, can become a HBC-related protein.

This work can help in oncology proteomics or serve as a model for other studies. In addition, S2SNet application is demonstrating his capacity to transform simple protein sequences in TIs and to be the base of numerous protein studies.

Acknowledgments

Cristian R. Munteanu thanks the FCT (Portugal) for support from grant SFRH/BPD/24997/2005. González-Díaz H. acknowledges program Isidro Parga Pondal of Xunta de Galicia by financial support of a tenure-eligible research position at the Faculty of pharmacy, University of Santiago de Compostela (Spain).

References

Agüero-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.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., and Reusser, F., 1993a. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J Biol Chem 268, 6119-6124.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., and Reusser, F., 1993b. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6548-6554.

Althaus, I.W., Chou, J.J., Gonzales, A.J., LeMay, R.J., Deibel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., and et al., 1994. Steady-state kinetic studies with the polysulfonate U-9843, an HIV reverse transcriptase inhibitor. Experientia 50, 23-8.

Althaus, I.W., Chou, K.C., Lemay, R.J., Franks, K.M., Deibel, M.R., Kezdy, F.J., Resnick, L., Busso, M.E., So, A.G., Downey, K.M., Romero, D.L., Thomas, R.C., Aristoff, P.A., Tarpley, W.G., and Reusser, F., 1996. The benzylthio-

pyrimidine U-31,355, a potent inhibitor of HIV-1 reverse transcriptase. Biochem Pharmacol 51, 743-50.

Anwar, R., Miloszewski, K.J., and Markham, A.F., 1998. New splicing mutations in the human factor XIIIA gene, each producing multiple mutant transcripts of varying abundance. Thromb Haemost 79, 1151-6.

Bielinska-Waz, D., Nowak, W., Waz, P., Nandy, A., and Clark, T., 2007. Distribution Moments of 2D-graphs as Descriptors of DNA Sequences. Chem. Phys. Lett. 443, 408-413.

Castillo-Garit, J.A., Marrero-Ponce, Y., Torrens, F., Garcia-Domenech, R., and Romero-Zaldivar, V., 2008. Bond-based 3D-chiral linear indices: Theory and QSAR applications to central chirality codification. J Comput Chem. 29(15), 2500-12.

Costantini, V., Zacharski, L.R., Memoli, V.A., Kisiel, W., Kudryk, B.J., and Rousseau, S.M., 1991. Fibrinogen deposition without thrombin generation in primary human breast cancer tissue. Cancer Res 51, 349-53.

Chen, Y.L., and Li, Q.Z., 2007a. Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245, 775-83.

Chen, Y.L., and Li, Q.Z., 2007b. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J Theor Biol 248, 377-381.

Chou, K.C., 1989. Graphical rules in steady and non-steady enzyme kinetics J Biol Chem 264, 12074-12079.

Chou, K.C., 1990. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems Biophys Chem 35, 1-24.

Chou, K.C., and Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. Biochem J 187, 829-835.

Chou, K.C., and Liu, W.M., 1981. Graphical rules for non-steady state enzyme kinetics. J Theor Biol 91, 637-54.

Chou, K.C., and Zhang, C.T., 1992. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Research and Human Retroviruses 8, 1967-1976.

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

Chou, K.C., and Shen, H.B., 2007. Recent progress in protein subcellular location prediction. Analytical Biochemistry 370, 1-16.

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

Chou, K.C., Kezdy, F.J., and Reusser, F., 1994. Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal Biochem 221, 217-230.

Chou, K.C., Zhang, C.T., and Elrod, D.W., 1996. Do "antisense proteins" exist? J Protein Chem 15, 59-61.

Dayhoff, M.O., A Model of Evolutionary Change, in: Dayhoff, M. O., (Ed.), Proteins in Atlas of Protein Sequence and Structure, Vol. 5 Supplement 3. Georgetown University Medical Center, National Biomedical Research Foundation 1978, pp. 345-358.

Devillers, J., and Balaban, A.T., 1999. Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach, The Netherlands.

Diao, Y., Li, M., Feng, Z., Yin, J., and Pan, Y., 2007. The community structure of human cellular signaling network. J Theor Biol 247, 608-15.

Ding, Y.S., Zhang, T.L., and Chou, K.C., 2007. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept Lett 14, 811-5.

Dobson, P.D., and Doig, A.J., 2005. Predicting enzyme class from protein structure without alignments. J Mol Biol 345, 187-99.

Dobson, P.D., Cai, Y.D., Stapley, B.J., and Doig, A.J., 2004. Prediction of protein function in the absence of significant sequence similarity. Curr Med Chem 11, 2135-42.

Estrada, E., and Molina, E., 2001. 3D connectivity indices in QSPR/QSAR studies. J Chem Inf Comput Sci 41, 791-7.

Ferino, G., Gonzalez-Diaz, H., Delogu, G., Podda, G., and Uriarte, E., 2008. Using spectral moments of spiral networks based on PSA/mass spectra outcomes to derive quantitative proteome-disease relationships (QPDRs) and predicting prostate cancer. Biochem Biophys Res Commun, doi:10.1016/j.bbrc.2008.05.071.

Fox, B.A., Yee, V.C., Pedersen, L.C., Le Trong, I., Bishop, P.D., Stenkamp, R.E., and Teller, D.C., 1999. Identification of the calcium binding site and a novel ytterbium site in blood coagulation factor XIII by x-ray crystallography. J Biol Chem 274, 4917-23.

Gao, L., Ding, Y.S., Dai, H., Shao, S.H., Huang, Z.D., and Chou, K.C., 2006. A novel fingerprint map for detecting SARS-CoV. Journal of Pharmaceutical and Biomedical Analysis 41, 246-50.

Gonzalez-Diaz, H., Gonzalez-Diaz, Y., Santana, L., Ubeira, F.M., and Uriarte, E., 2008. Proteomics, networks and connectivity indices. Proteomics 8, 750-78.

González-Díaz, H., Sanchez-Gonzalez, A., and Gonzalez-Diaz, Y., 2006. 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. J Inorg Biochem 100, 1290-7.

González-Díaz, H., Vilar, S., Santana, L., and Uriarte, E., 2007a. Medicinal Chemistry and Bioinformatics – Current Trends in Drugs Discovery with Networks Topological Indices. Curr Top Med Chem 7, 1025-39.

González-Díaz, H., Ferino, G., Podda, G., and Uriarte, E., 2007b. Discriminating Prostate Cancer Patients from control group with connectivity indices. ECSOC 11, G1:1-10.

González-Díaz, H., Bonet, I., Terán, C., de Clercq, E., Bello, R., García, M., Santana, L., and Uriarte, E., 2007c. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. European Journal of Medicinal Chemistry 42, 580-585.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., and Thun, M.J., 2008. Cancer statistics, 2008. CA Cancer J Clin 58, 71-96.

Jiang, X., Wei, R., Zhang, T., and Gu, Q., 2008. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept Lett 15, 392-6.

Koutsofios, E., and North, S.C., 1993. Drawing Graphs with dot. AT&T Bell Laboratories, Murray Hill, NJ, USA.

Kowalski, R.D., and Wold, S., Pattern recognition in chemistry, in: Krishnaiah, P. R. and Kanal, L. N., Eds.), Handbook of Statistic, North Holland Publishing Company, Amsterdam 1982, pp. 673-697.

Kuzmic, P., Ng, K.Y., and Heath, T.D., 1992. Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation Anal Biochem 200 68-73.

Li, F.M., and Li, Q.Z., 2008. Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34, 119-25.

Liao, B., and Wang, T.M., 2004. Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. J Chem Inf Comput Sci 44, 1666-70.

Liao, B., and Ding, K., 2005. Graphical approach to analyzing DNA sequences. J Comput Chem 26, 1519-23.

Lin, H., 2008. The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252, 350-6.

Marrero-Ponce, Y., Diaz, H.G., Zaldivar, V.R., Torrens, F., and Castro, E.A., 2004. 3D-chiral quadratic indices of the 'molecular pseudograph's atom adjacency matrix' and their application to central chirality codification: classification of ACE inhibitors and prediction of sigma-receptor antagonist activities. Bioorg Med Chem 12, 5331-42.

Morales, A.H., Cabrera Perez, M.A., and Gonzalez, M.P., 2006. A radial-distribution-function approach for predicting rodent carcinogenicity. J Mol Model 12, 769-80.

Munteanu CR and Gonzáles-Diáz H., 2008. S2SNet - Sequence to Star Network, Reg. No. 03 / 2008 / 1338, Santiago de Compostela, Spain.

Niu, B., Cai, Y.D., Lu, W.C., Li, G.Z., and Chou, K.C., 2006. Predicting protein structural class with AdaBoost Learner. Protein Pept Lett 13, 489-92.

Prado-Prado, F.J., González-Díaz, H., Martinez de la Vega, O., Ubeira, F.M., and Chou, K.C., 2008. Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. Bioorganic and Medicinal Chemistry 16, 5871–5880.

Qi, X.Q., Wen, J., and Qi, Z.H., 2007. New 3D graphical representation of DNA sequence based on dual nucleotides. J Theor Biol 249, 681-90.

Randic, M., 2000. Condensed representation of DNA primary sequences. J Chem Inf Comput Sci 40, 50-6.

Randic, M., Vracko, M., Nandy, A., and Basak, S.C., 2000. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. . J Chem Inf Comput Sci 40, 1235-1244.

Randic, M., and Basak, S.C., 2001. Characterization of DNA primary sequences based on the average distances between bases. J. Chem. Inf. Comput. Sci. 41, 561-8.

Randic, M., and Balaban, A.T., 2003. On a four-dimensional representation of DNA primary sequences. J Chem Inf Comput Sci 43, 532-9.

Randic, M., Zupan, J., and Vikic-Topic, D., 2007. On representation of proteins by star-like graphs. J Mol Graph Model, 290-305.

Rappin, N., and Dunn, R., 2006. wxPython in Action. Manning Publications Co., Greenwich, CT.

Sjoblom, 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, 268-74.

StatSoft.Inc., STATISTICA (data analysis software system), version 6.0, www.statsoft.com.Statsoft, 2002.

Takahashi, H., Isobe, T., Horibe, S., Takagi, J., Yokosaki, Y., Sheppard, D., and Saito, Y., 2000. Tissue transglutaminase, coagulation factor XIII, and the pro-polypeptide of von Willebrand factor are all ligands for the integrins alpha 9beta 1 and alpha 4beta 1. J Biol Chem 275, 23589-95.

Todeschini, R., and Consonni, V., 2002. Handbook of Molecular Descriptors. Wiley-VCH.

Van Waterbeemd, H., Discriminant Analysis for Activity Prediction, in: Van Waterbeemd, H., (Ed.), Chemometric methods in molecular design, Vol. 2. Wiley-VCH, New York 1995, pp. 265-282.

Vilar, S., Gonzalez-Diaz, H., Santana, L., and Uriarte, E., 2008. QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks. J Comput Chem.

Wang, M., Yao, J.S., Huang, Z.D., Xu, Z.J., Liu, G.P., Zhao, H.Y., Wang, X.Y., Yang, J., Zhu, Y.S., and Chou, K.C., 2005. A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis. Med Chem 1, 39-47.

WHO, Cancer, World Health Organization, Fact sheet N°297, http://www.who.int/mediacentre/factsheets/fs297/en/ 2008.

Wolfram, S., 1984. Cellular automation as models of complexity. Nature 311, 419-424.

Wolfram, S., 2002. A New Kind of Science. Wolfram Media Inc., Champaign, IL.

Xiao, X., and Chou, K.C., 2007. Digital coding of amino acids based on hydrophobic index. Protein Pept Lett 14, 871-5.

Xiao, X., Shao, S.H., and Chou, K.C., 2006a. A probability cellular automaton model for hepatitis B viral infections. Biochemical and Biophysical Research Communications 342, 605-10.

Xiao, X., Shao, S., Ding, Y., Huang, Z., and Chou, K.C., 2006b. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30, 49-54.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., and Chou, K.C., 2005a. Using cellular automata to generate image representation for biological sequences. Amino Acids 28, 29-35.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., and Chou, K.C., 2005b. An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J Theor Biol 235, 555-65.

Yee, V.C., Pedersen, L.C., Le Trong, I., Bishop, P.D., Stenkamp, R.E., and Teller, D.C., 1994. Three-dimensional structure of a transglutaminase: human blood coagulation factor XIII. Proc Natl Acad Sci U S A 91, 7296-300.

Zhang, C.T., and Chou, K.C., 1994. Analysis of codon usage in 1562 E. Coli protein coding sequences. J Mol Biol 238, 1-8.

Zhang, T.L., Ding, Y.S., and Chou, K.C., 2008. Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. J Theor Biol 250, 186-93.

Zhou, X.B., Chen, C., Li, Z.C., and Zou, X.Y., 2007. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248, 546-51.

LEGEND FOR FIGURE

Figure 1a. The non-embedded star graphs for PRPS1.

Figure 1b. The embedded star graphs for PRPS1.

Figure 2. Graphical representation of two-way joining cluster analysis of the HBC probability after the mutations.

Figure 3. Graphical representation of reduced values of the reordered data matrix by tw-JCA method for HBC probability after the mutations.

Figure 4. Graphical representation of two-way joining cluster analysis of the HCC probability after the mutations.

Figure 5. Graphical representation of reduced values of the reordered data matrix by tw-JCA method for HCC probability after the mutations.

Figure 6. Graphical representation of two-way joining cluster analysis of the probability of the mutated HBC-related proteins to turn into non-cancer proteins.

Figure 7. Graphical representation of two-way joining cluster analysis of the probability of the mutated HCC-related proteins to turn into non-cancer proteins.

LEGEND FOR TABLES

Table 1. Single amino acid mutations and the corresponding log-odd PAM250 value.

Table 2. Training/predicting accuracies of Cancer (*C*) / non-cancer (*nC*) models using embedded (*E*) and non-embedded (*nE*) Star Graph TIs, **p**TIs and **d**TIs.

Table 3. Accuracy of input-coded multi-target and individual HBC and HCC classification models based on the embedded TIs (TI**e**+**d**TI**e** and **p**TI**e**).

Table 1:

Original AA	Mutated AA	log-odd PAM250	Notation
D	N	2	D->N / 2DN
E	Q	2	E->Q / 2EQ
F	L	2	F->L / 2FL
H	N	2	H->N / 2HN
H	R	2	H->R / 2HR
L	I	2	L->I / 2LI
M	I	2	M->I / 2MI
Q	D	2	Q->D / 2QD
V	L	2	V->L / 2VL
V	M	2	V->M / 2VM
W	R	2	W->R / 2WR
E	D	3	E->D / 3ED
H	Q	3	H->Q / 3HQ
K	R	3	K->R / 3KR
M	L	4	M->L / 4ML
V	I	4	V->I / 4VI
Y	F	7	Y->F / 7YF

Table 2:

Star Graph Type	Attributes	Train			Cross-validation			Total			Eq. Vars.
		nC (%)	*C* (%)	*Total* (%)	*nC* (%)	*C* (%)	*Total* (%)	*nC* (%)	*C* (%)	*Total* (%)	
nE	**p**TI	90.4	69.4	88.5	91.4	66.0	89.1	90.7	68.6	88.7	4
	TI, **p**TI	90.4	68.1	88.3	90.8	66.0	88.5	90.5	67.5	88.4	5
	dTI	86.0	79.9	85.4	87.0	74.5	85.9	86.2	78.5	85.5	2
	TI, **d**TI	88.1	74.3	86.9	88.9	72.3	87.4	88.3	73.8	87.0	4
	TI, **p**TI, **d**TI	91.1	66.0	88.8	91.8	61.7	89.1	91.3	64.9	88.9	6
E	**p**TI**e**	92.3	70.1	**90.3**	93.1	70.2	**91.0**	92.5	70.2	**90.5**	**6**
	TI**e**, **p**TI**e**	92.7	69.4	90.6	93.3	70.2	91.2	92.8	69.6	90.7	6
	dTI**e**	88.1	78.5	87.3	88.3	76.6	87.2	88.2	78.0	87.3	4
	TI**e**, **d**TI**e**	91.4	75.7	**89.9**	91.8	74.5	**90.3**	91.5	75.4	**90.0**	**5**
	TI**e**, **p**TI**e**, **d**TI**e**	93.1	68.1	90.8	93.3	66.0	90.8	93.1	67.5	90.8	8
nE* & *E	**p**TI, **p**TI**e**	90.2	70.1	88.4	91.2	68.1	89.1	90.5	69.6	88.6	4
	TI, TI**e**, **p**TI, **p**TI**e**	92.3	68.8	90.1	92.0	66.0	89.7	92.2	68.1	90.0	8
	dTI, **d**TI**e**	90.3	78.5	89.2	90.6	76.6	89.3	90.4	78.0	89.2	6
	TI, TI**e**, **d**TI, **d**TI**e**	90.9	72.9	89.3	91.4	72.3	89.7	91.1	72.8	89.4	7
	TI, TI**e**, **p**TI, **p**TI**e**, **d**TI, **d**TI**e**	92.3	68.8	90.1	92.2	70.2	90.3	92.3	69.1	90.2	8

Table 3:

Eq.	Cancer	Correct	Incorrect	Accuracy
		Tl**e, d**T**l**e**		
14	**Both**	307	1795	90.0%
14a	**HBC**	168	880	91.8%
14b	**HCC**	139	915	88.2%
		pT**l**e**		
15	**Both**	277	1825	90.5%
15a	**HBC**	170	878	91.8%
15b	**HCC**	107	947	89.2%

Figure 1a:

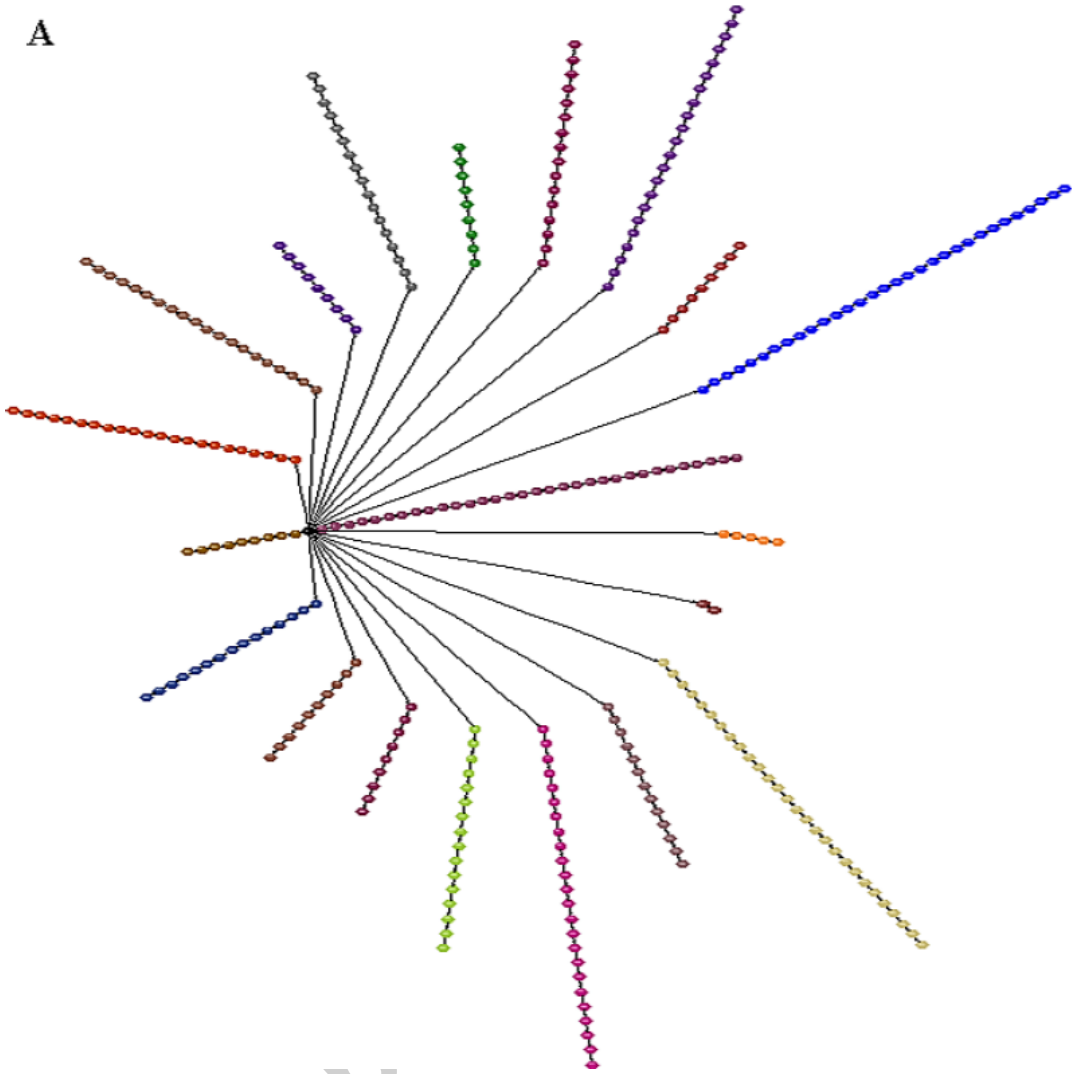

Figure 1b:

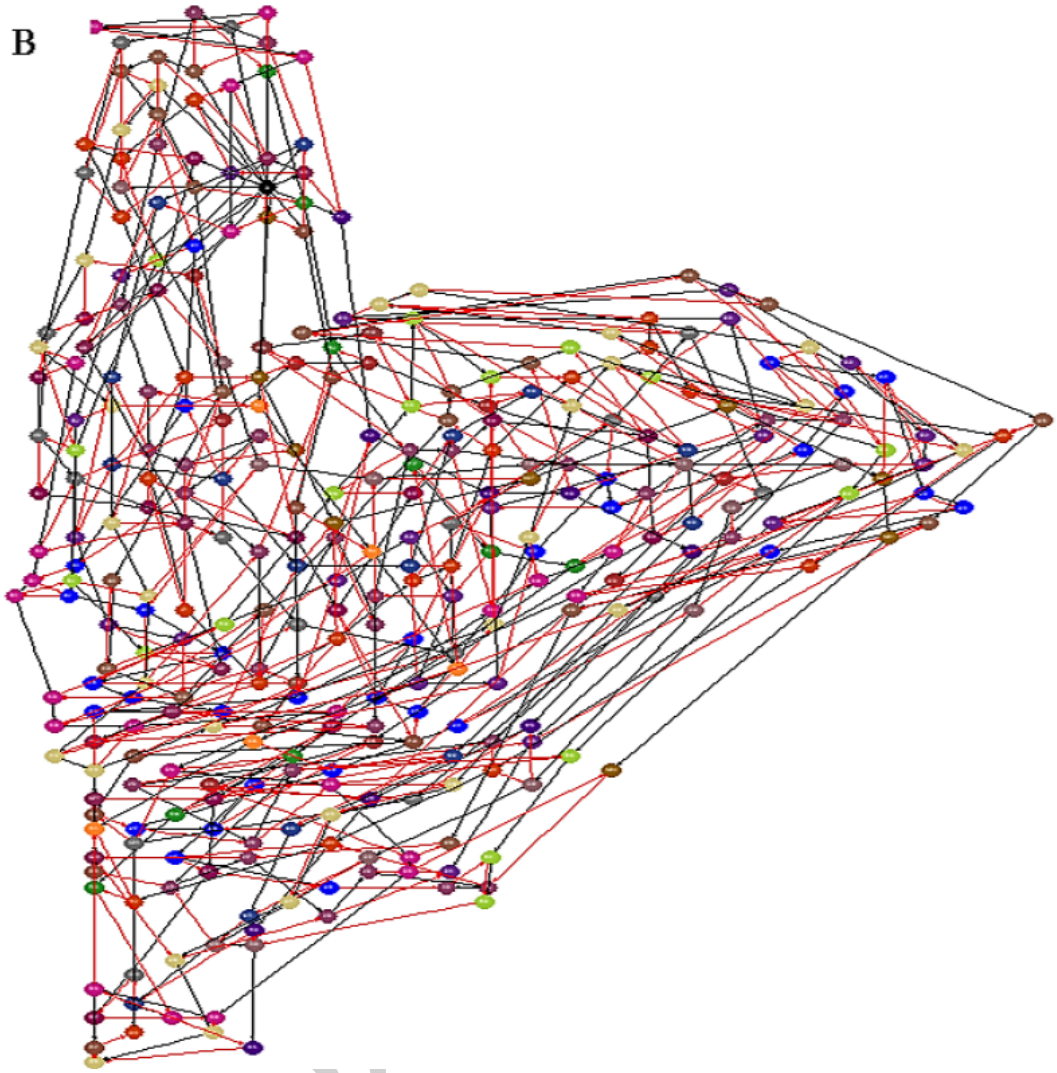

Figure 2:

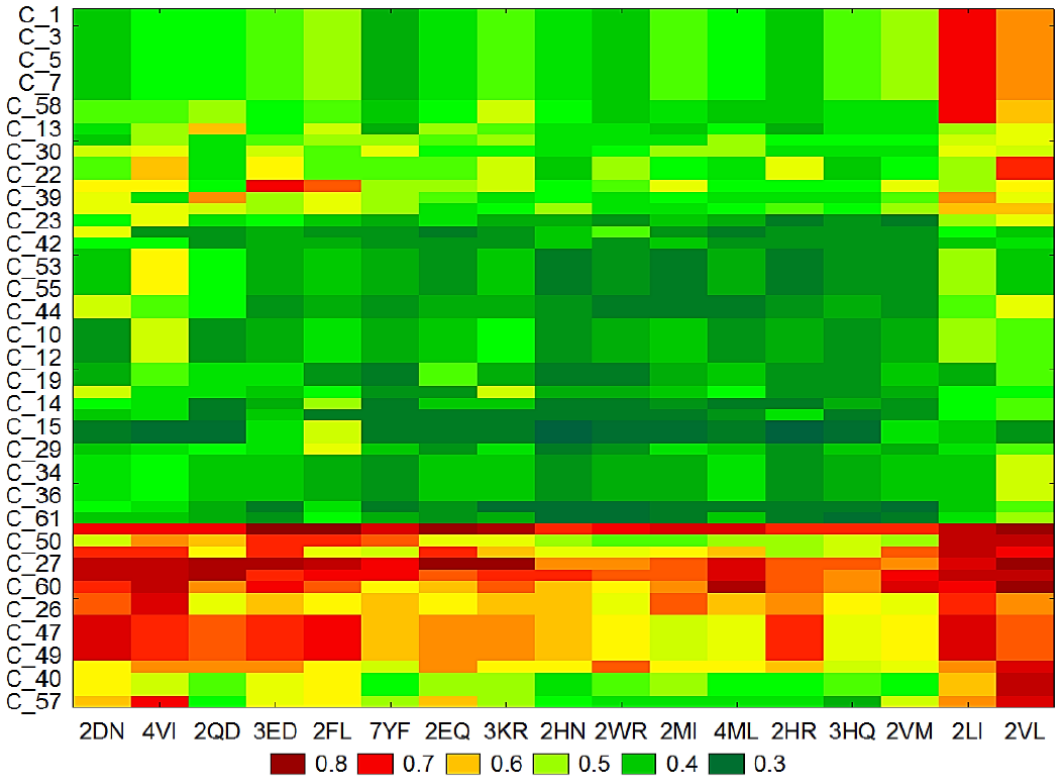

Figure 3:

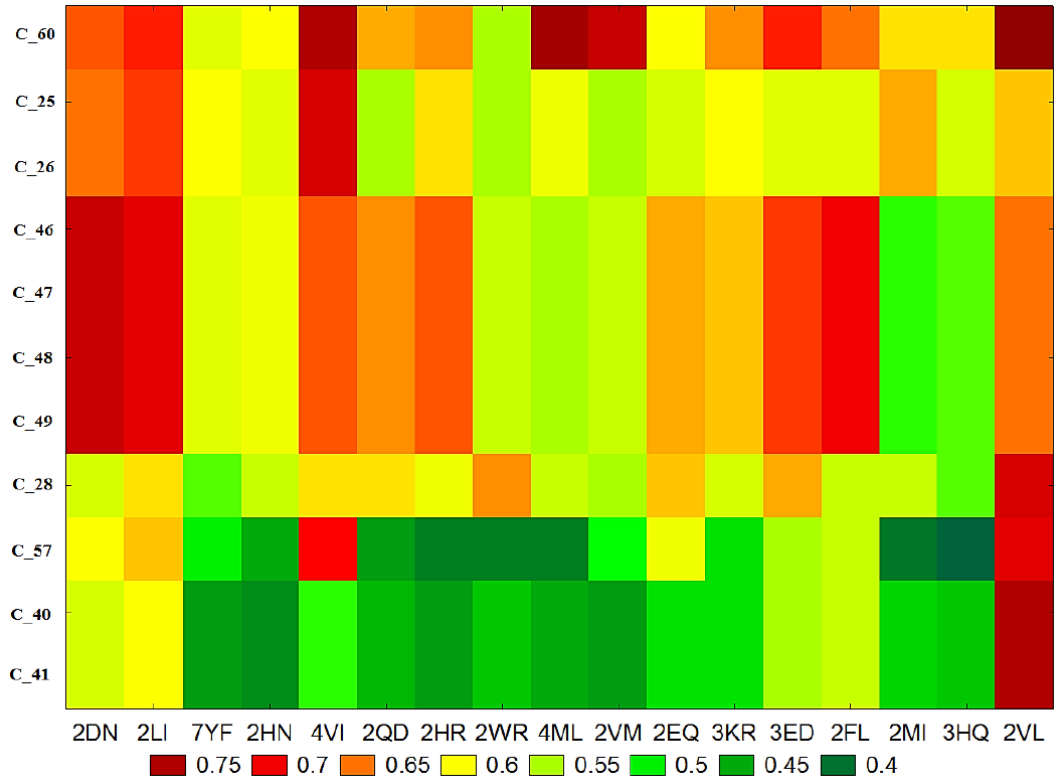

Figure 4:

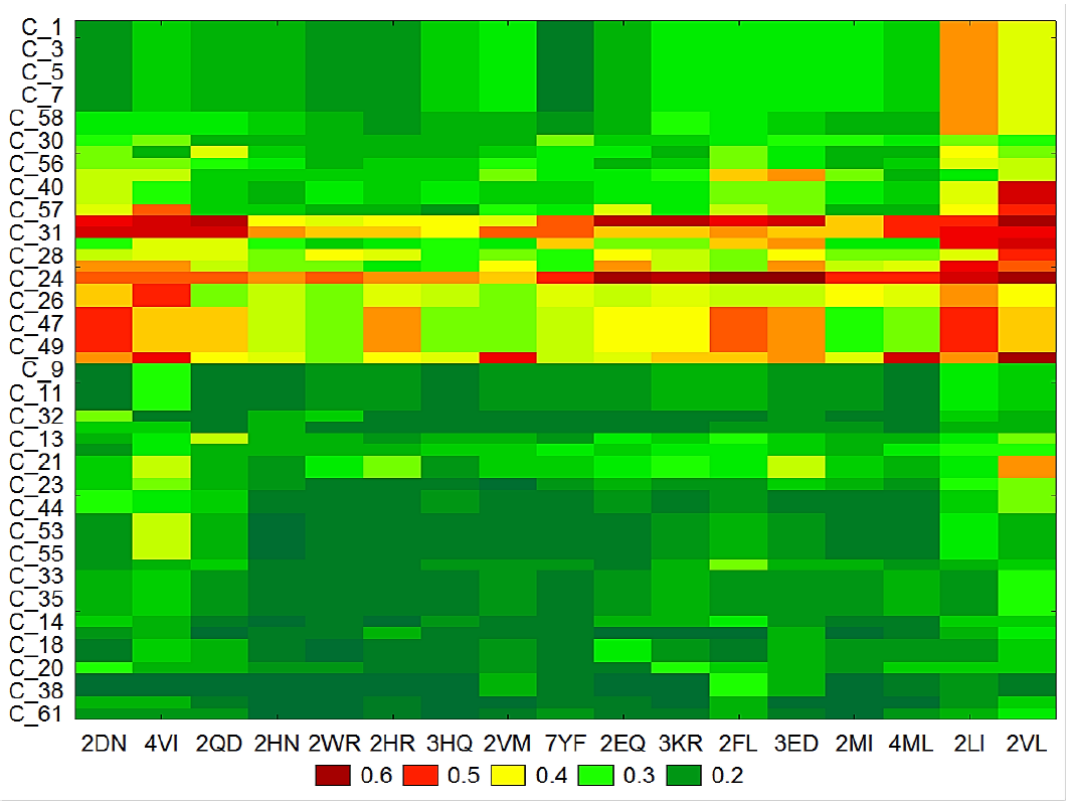

Figure 5:

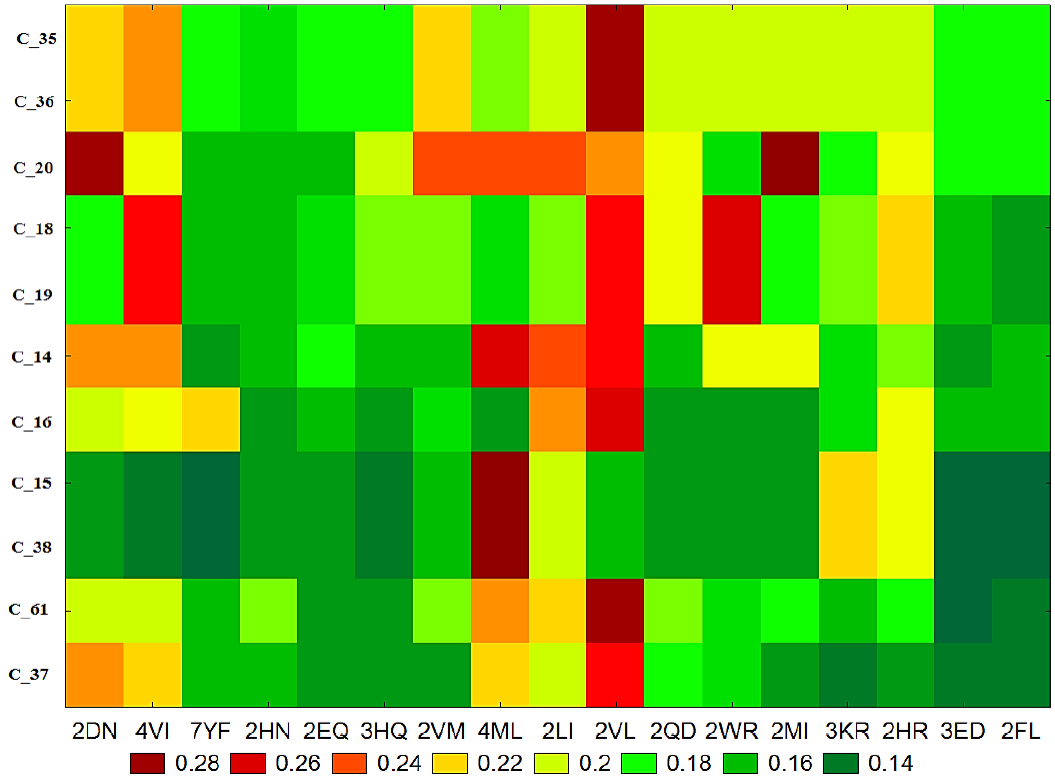

Figure 6:

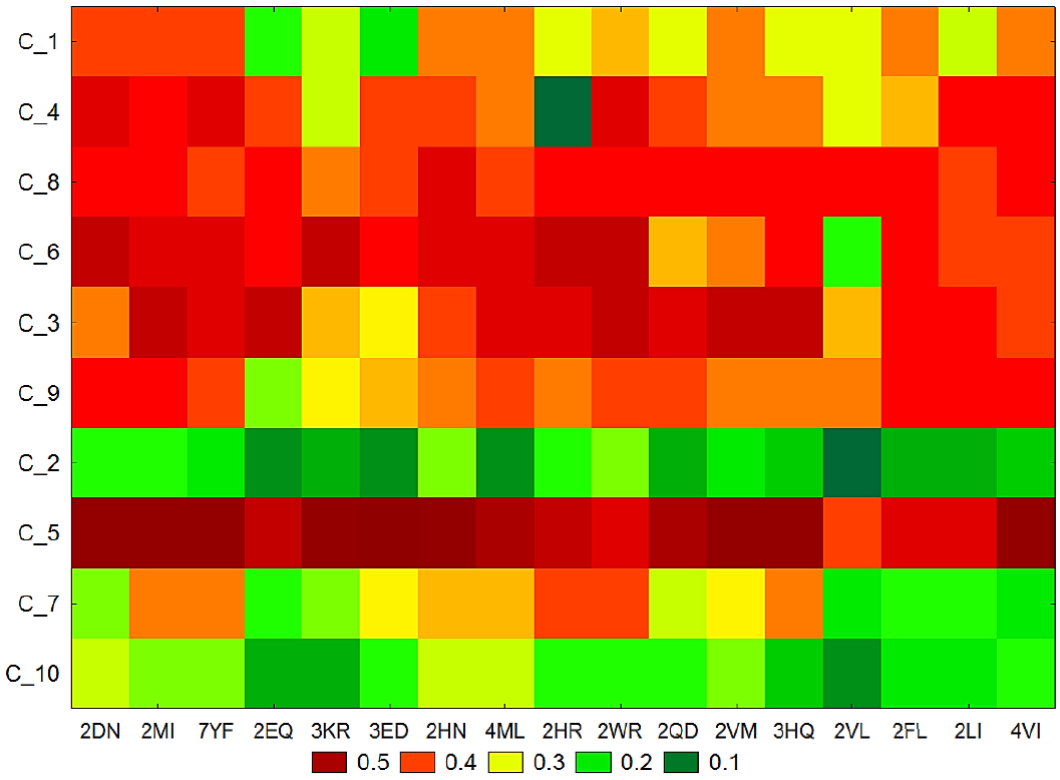

Figure 7:

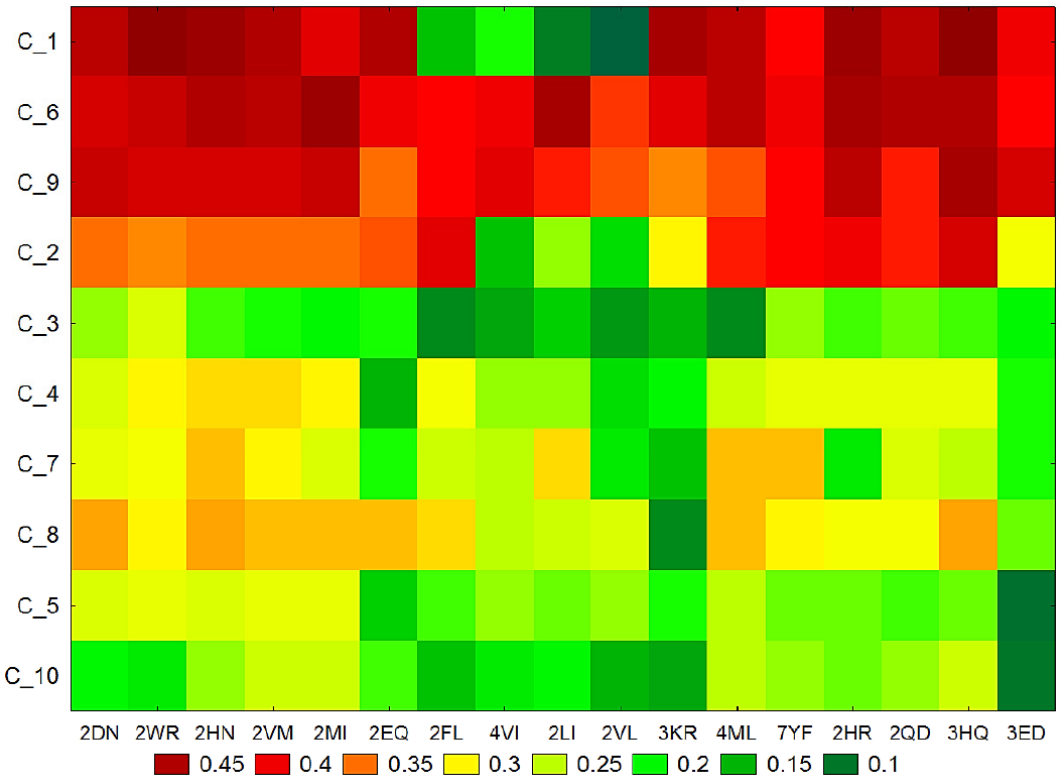

Table S1. Training/cross-validation selection test

	CV0	CV1	CV2	CV3	CV4	CV5	CV6	CV7	CV8	CV9	Average
Total **Train** (%)	89.9	89.3	90.1	88.5	90.4	89.8	91.6	90.1	92.7	89.4	**90.2**
Total **CV** (%)	90.3	89.5	90.6	89.3	89.1	89.7	87.4	90.6	89.3	90.1	**89.6**

Table S2. The equations of the Cancer (*C*) / non-cancer (*nC*) models based on the embedded (E) and non-embedded (nE) Star Graph TIs, **p**TIs and **d**TIs

Star Graph Type	Attributes	Equation of the QPDR models
n**E**	**p**TI	-4.7-9.1*pTr4+16.7*pH-13.9*pW+8.8*pJ
	TI, **p**TI	-4.7-1.6*pW-44.4*pX4+49.6*pX5-6.0*W+4.8*J
	dTI	-3.8-12.5*dS6+14.4*dX5
	TI, **d**TI	-4.2-14.1*Tr4-6.7*dS+4.2*J+18.7*X5
	TI, **p**TI, **d**TI	-4.7-1.5*pW+5.2*pX5+335.1*dH-11.0*W-333.0*dS6+7.7*J
E	**p**TI**e**	-4.1-118.6*pTr0e+80.7*pTr2e+1.4*pTr3e+100.3*pSe-101.4*pJe+39.7*pX2e
	TI**e**, **p**TI**e**	-4.2+2.8*pTr3e-1.6*Tr5e+45.0*He-42.9*S6e+111.9*Se-113.2*Je
	dTI**e**	-3.1+9.8*dTr2e+0.6*dTr3e-8.6*dS6e-0.1*dJe
	TI**e**, **d**TI**e**	-4.4+1.7*Tr3e+124.8*Se-126.5*dJe+48.6*dX2e-45.9*X5e
	TI**e**, **p**TI**e**, **d**TI**e**	-4.2+4.1*pTr3e+5.5*pTr4e-5.8*pS6e-3.4*dTr5e+35.5*He-33.7*dS6e+96.7*dSe-97.1*Je
n**E** & **E**	**p**TI, **p**TI**e**	-4.7+14.3*pX5+2.3*pTr3e-13.7*pS6e-0.3*pJe
	TI, TI**e**, **p**TI, **p**TI**e**	-5.0+28.0*pX5+1.6*pTr3e+7.8*pTr4e+123.3*pSe-124.5*pJe-32.1*pX5e-5.2*W+4.0*J
	dTI, **d**TI**e**	-4.1-154.2*dW+2.8*dJ+20.9*dTr4e+132.2*dSe+17.4*dJe-16.7*dX5e
	TI, TI**e**, **d**TI, **d**TI**e**	-4.4+11.9*dH-82.5*dS+6.3*dJ+1.1*Tr3e+17.8*dTr4e-23.9*X3e
	TI, TI**e**, **p**TI, **p**TI**e**, **d**TI, **d**TI**e**	-4.9+1.8*pTr3e+21.0*pTr4e-pJe-17.6*pX4e-17.5*W-81.2*dS+2.8*dJ+94.8*dSe

Table S3. tw-JCA reordered data matrix for the probability of the mutated non-cancer proteins to turn into HBC-related proteins

Protein	Name	2DN	4VI	2QD	3ED	2FL	7YF	2EQ	3KR	2HN	2WR	2MI	4ML	2HR	3HQ	2VM	2LI	2VL
1A49 A	C_1	0.390	0.432	0.425	0.464	0.490	0.355	0.402	0.470	0.407	0.376	0.466	0.446	0.390	0.460	0.483	0.680	0.602
1A49 B	C_2	0.390	0.432	0.425	0.464	0.490	0.355	0.402	0.470	0.407	0.376	0.466	0.446	0.390	0.460	0.483	0.680	0.602
1A49 C	C_3	0.390	0.432	0.425	0.464	0.490	0.355	0.402	0.470	0.407	0.376	0.466	0.446	0.390	0.460	0.483	0.680	0.602
1A49 D	C_4	0.390	0.432	0.425	0.464	0.490	0.355	0.402	0.470	0.407	0.376	0.466	0.446	0.390	0.460	0.483	0.680	0.602
1A49 E	C_5	0.390	0.432	0.425	0.464	0.490	0.355	0.402	0.470	0.407	0.376	0.466	0.446	0.390	0.460	0.483	0.680	0.602
1A49 F	C_6	0.390	0.432	0.425	0.464	0.490	0.355	0.402	0.470	0.407	0.376	0.466	0.446	0.390	0.460	0.483	0.680	0.602
1A49 G	C_7	0.390	0.432	0.425	0.464	0.490	0.355	0.402	0.470	0.407	0.376	0.466	0.446	0.390	0.460	0.483	0.680	0.602
1A49 H	C_8	0.390	0.432	0.425	0.464	0.490	0.355	0.402	0.470	0.407	0.376	0.466	0.446	0.390	0.460	0.483	0.680	0.602
2FOK A	C_58	0.470	0.468	0.476	0.449	0.470	0.379	0.429	0.513	0.446	0.394	0.402	0.396	0.380	0.407	0.401	0.679	0.588
2FOK B	C_59	0.470	0.468	0.476	0.449	0.470	0.379	0.429	0.513	0.446	0.394	0.402	0.396	0.380	0.407	0.401	0.679	0.588
1AI4 B	C_13	0.409	0.488	0.577	0.443	0.503	0.367	0.476	0.455	0.408	0.403	0.398	0.429	0.371	0.415	0.423	0.479	0.531
1QBQ B	C_45	0.385	0.490	0.419	0.458	0.484	0.493	0.461	0.475	0.409	0.402	0.409	0.482	0.425	0.447	0.438	0.509	0.525
1F83 A	C_30	0.516	0.542	0.415	0.515	0.469	0.529	0.436	0.439	0.421	0.429	0.499	0.484	0.400	0.413	0.410	0.544	0.519
1DJN A	C_21	0.450	0.580	0.409	0.556	0.471	0.454	0.467	0.509	0.383	0.493	0.432	0.414	0.528	0.392	0.431	0.490	0.664
1DJN B	C_22	0.450	0.580	0.409	0.556	0.471	0.454	0.467	0.509	0.383	0.493	0.432	0.414	0.528	0.392	0.431	0.490	0.664
1C4A A	C_17	0.561	0.570	0.445	0.676	0.637	0.483	0.479	0.525	0.447	0.457	0.536	0.425	0.437	0.447	0.534	0.493	0.567
1KAP P	C_39	0.541	0.416	0.601	0.478	0.529	0.480	0.471	0.423	0.432	0.411	0.426	0.423	0.413	0.407	0.451	0.610	0.537
1QQ1 A	C_56	0.530	0.531	0.500	0.481	0.541	0.479	0.414	0.434	0.479	0.415	0.418	0.445	0.458	0.445	0.499	0.585	0.583
1DLJ A	C_23	0.445	0.542	0.415	0.436	0.380	0.374	0.401	0.371	0.357	0.349	0.396	0.368	0.317	0.342	0.313	0.496	0.538
1JDA	C_32	0.527	0.332	0.346	0.357	0.325	0.339	0.315	0.334	0.396	0.461	0.328	0.316	0.330	0.341	0.328	0.430	0.400
1PLU A	C_42	0.443	0.446	0.343	0.373	0.357	0.335	0.335	0.336	0.396	0.336	0.379	0.334	0.340	0.340	0.334	0.398	0.412
1QMG A	C_52	0.393	0.561	0.429	0.364	0.398	0.352	0.336	0.387	0.307	0.333	0.318	0.352	0.321	0.342	0.327	0.487	0.398
1QMG B	C_53	0.393	0.561	0.429	0.364	0.398	0.352	0.336	0.387	0.307	0.333	0.318	0.352	0.321	0.342	0.327	0.487	0.398
1QMG C	C_54	0.393	0.561	0.429	0.364	0.398	0.352	0.336	0.387	0.307	0.333	0.318	0.352	0.321	0.342	0.327	0.487	0.398
1QMG D	C_55	0.393	0.561	0.429	0.364	0.398	0.352	0.336	0.387	0.307	0.333	0.318	0.352	0.321	0.342	0.327	0.487	0.398
1QBI A	C_43	0.501	0.465	0.445	0.347	0.362	0.347	0.368	0.351	0.334	0.320	0.320	0.320	0.347	0.373	0.332	0.453	0.528
1QBI B	C_44	0.501	0.465	0.445	0.347	0.362	0.347	0.368	0.351	0.334	0.320	0.320	0.320	0.347	0.373	0.332	0.453	0.528
1ADO A	C_9	0.349	0.503	0.344	0.362	0.402	0.356	0.383	0.427	0.346	0.369	0.380	0.342	0.367	0.341	0.367	0.483	0.451
1ADO B	C_10	0.349	0.503	0.344	0.362	0.402	0.356	0.383	0.427	0.346	0.369	0.380	0.342	0.367	0.341	0.367	0.483	0.451
1ADO C	C_11	0.349	0.503	0.344	0.362	0.402	0.356	0.383	0.427	0.346	0.369	0.380	0.342	0.367	0.341	0.367	0.483	0.451
1ADO D	C_12	0.349	0.503	0.344	0.362	0.402	0.356	0.383	0.427	0.346	0.369	0.380	0.342	0.367	0.341	0.367	0.483	0.451
1CCW B	C_18	0.350	0.453	0.404	0.422	0.343	0.319	0.466	0.361	0.315	0.307	0.369	0.376	0.330	0.339	0.379	0.366	0.456

1CCW D	C_19	0.350	0.453	0.404	0.422	0.343	0.319	0.466	0.361	0.315	0.307	0.369	0.376	0.330	0.339	0.379	0.366	0.456
1D5T A	C_20	0.500	0.400	0.408	0.398	0.441	0.329	0.338	0.508	0.356	0.355	0.381	0.438	0.328	0.327	0.358	0.447	0.431
1AQ2	C_14	0.434	0.424	0.325	0.369	0.478	0.316	0.397	0.396	0.307	0.321	0.327	0.322	0.311	0.356	0.343	0.439	0.458
1C1D B	C_16	0.383	0.401	0.311	0.394	0.309	0.315	0.311	0.313	0.319	0.323	0.307	0.343	0.409	0.318	0.333	0.428	0.469
1BY7 A	C_15	0.305	0.288	0.298	0.404	0.510	0.315	0.302	0.308	0.271	0.277	0.292	0.322	0.273	0.300	0.421	0.386	0.331
1JRR A	C_38	0.305	0.288	0.298	0.404	0.510	0.315	0.302	0.308	0.271	0.277	0.292	0.322	0.273	0.300	0.421	0.386	0.331
1F6W A	C_29	0.393	0.425	0.444	0.406	0.541	0.384	0.349	0.401	0.325	0.333	0.403	0.380	0.346	0.389	0.368	0.410	0.454
1JDF A	C_33	0.412	0.431	0.384	0.383	0.374	0.348	0.388	0.394	0.350	0.351	0.357	0.416	0.350	0.352	0.393	0.388	0.501
1JDF B	C_34	0.412	0.431	0.384	0.383	0.374	0.348	0.388	0.394	0.350	0.351	0.357	0.416	0.350	0.352	0.393	0.388	0.501
1JDF C	C_35	0.412	0.431	0.384	0.383	0.374	0.348	0.388	0.394	0.350	0.351	0.357	0.416	0.350	0.352	0.393	0.388	0.501
1JDF D	C_36	0.412	0.431	0.384	0.383	0.374	0.348	0.388	0.394	0.350	0.351	0.357	0.416	0.350	0.352	0.393	0.388	0.501
1JET A	C_37	0.425	0.411	0.351	0.305	0.423	0.324	0.337	0.300	0.282	0.297	0.313	0.313	0.322	0.310	0.290	0.388	0.457
3GCB	C_61	0.387	0.390	0.364	0.349	0.426	0.365	0.346	0.352	0.279	0.296	0.305	0.377	0.319	0.299	0.325	0.421	0.492
1DOT	C_24	0.693	0.687	0.698	0.812	0.801	0.709	0.795	0.765	0.672	0.689	0.712	0.717	0.668	0.668	0.655	0.745	0.788
1QFS A	C_50	0.509	0.609	0.588	0.666	0.657	0.637	0.533	0.546	0.496	0.452	0.465	0.478	0.487	0.501	0.491	0.737	0.747
1QME A	C_51	0.667	0.662	0.556	0.675	0.538	0.509	0.664	0.576	0.541	0.545	0.556	0.592	0.496	0.519	0.627	0.740	0.682
1EGU A	C_27	0.735	0.749	0.763	0.759	0.729	0.682	0.776	0.775	0.624	0.605	0.646	0.716	0.632	0.625	0.608	0.703	0.787
1GOF	C_31	0.745	0.749	0.755	0.654	0.677	0.693	0.642	0.655	0.670	0.642	0.639	0.712	0.637	0.612	0.683	0.735	0.729

Table S4. Reduced values of the reordered data matrix by tw-JCA method for HBC probability

Protein	Name	2DN	4VI	2QD	3ED	2FL	7YF	2EQ	3KR	2HN	2WR	2MI	4ML	2HR	3HQ	2VM	2LI	2VL
3DMR	C_60	0.660	0.742	0.624	0.679	0.644	0.575	0.592	0.639	0.596	0.550	0.602	0.759	0.631	0.608	0.724	0.677	0.780
1E6Y B	C_25	0.641	0.713	0.549	0.579	0.574	0.595	0.563	0.597	0.578	0.545	0.626	0.588	0.603	0.561	0.546	0.668	0.611
1E6Y E	C_26	0.641	0.713	0.549	0.579	0.574	0.595	0.563	0.597	0.578	0.545	0.626	0.588	0.603	0.561	0.546	0.668	0.611
1QF7 A	C_46	0.722	0.654	0.636	0.668	0.695	0.575	0.621	0.611	0.581	0.555	0.512	0.544	0.658	0.527	0.551	0.708	0.649
1QF7 B	C_47	0.722	0.654	0.636	0.668	0.695	0.575	0.621	0.611	0.581	0.555	0.512	0.544	0.658	0.527	0.551	0.708	0.649
1QF7 C	C_48	0.722	0.654	0.636	0.668	0.695	0.575	0.621	0.611	0.581	0.555	0.512	0.544	0.658	0.527	0.551	0.708	0.649
1QF7 D	C_49	0.722	0.654	0.636	0.668	0.695	0.575	0.621	0.611	0.581	0.555	0.512	0.544	0.658	0.527	0.551	0.708	0.649
1EUU	C_28	0.568	0.605	0.601	0.623	0.555	0.525	0.611	0.562	0.553	0.633	0.555	0.555	0.588	0.522	0.541	0.603	0.716
1NGS A	C_40	0.564	0.513	0.460	0.542	0.553	0.436	0.487	0.485	0.423	0.468	0.478	0.447	0.434	0.468	0.436	0.596	0.743
1NGS B	C_41	0.564	0.513	0.460	0.542	0.553	0.436	0.487	0.485	0.423	0.468	0.478	0.447	0.434	0.468	0.436	0.596	0.743
1QRK B	C_57	0.593	0.688	0.437	0.548	0.559	0.492	0.583	0.481	0.442	0.417	0.404	0.420	0.420	0.375	0.510	0.620	0.708

Table S5. tw-JCA reordered data matrix for the probability of the mutated non-cancer proteins to turn into HCC-related proteins

Protein	Name	2DN	4VI	2QD	3ED	2FL	7YF	2EQ	3KR	2HN	2WR	2MI	4ML	2HR	3HQ	2VM	2LI	2VL
1A49 A	C_1	0.198	0.227	0.222	0.209	0.189	0.198	0.247	0.264	0.175	0.205	0.255	0.270	0.250	0.252	0.237	0.450	0.368
1A49 B	C_2	0.198	0.227	0.222	0.209	0.189	0.198	0.247	0.264	0.175	0.205	0.255	0.270	0.250	0.252	0.237	0.450	0.368
1A49 C	C_3	0.198	0.227	0.222	0.209	0.189	0.198	0.247	0.264	0.175	0.205	0.255	0.270	0.250	0.252	0.237	0.450	0.368
1A49 D	C_4	0.198	0.227	0.222	0.209	0.189	0.198	0.247	0.264	0.175	0.205	0.255	0.270	0.250	0.252	0.237	0.450	0.368
1A49 E	C_5	0.198	0.227	0.222	0.209	0.189	0.198	0.247	0.264	0.175	0.205	0.255	0.270	0.250	0.252	0.237	0.450	0.368
1A49 F	C_6	0.198	0.227	0.222	0.209	0.189	0.198	0.247	0.264	0.175	0.205	0.255	0.270	0.250	0.252	0.237	0.450	0.368
1A49 G	C_7	0.198	0.227	0.222	0.209	0.189	0.198	0.247	0.264	0.175	0.205	0.255	0.270	0.250	0.252	0.237	0.450	0.368
1A49 H	C_8	0.198	0.227	0.222	0.209	0.189	0.198	0.247	0.264	0.175	0.205	0.255	0.270	0.250	0.252	0.237	0.450	0.368
2FOK A	C_58	0.254	0.253	0.259	0.237	0.201	0.191	0.209	0.205	0.190	0.224	0.289	0.255	0.239	0.206	0.202	0.449	0.354
2FOK B	C_59	0.254	0.253	0.259	0.237	0.201	0.191	0.209	0.205	0.190	0.224	0.289	0.255	0.239	0.206	0.202	0.449	0.354
1F83 A	C_30	0.291	0.313	0.214	0.219	0.225	0.204	0.213	0.211	0.302	0.229	0.231	0.254	0.290	0.278	0.266	0.315	0.294
1KAP P	C_39	0.312	0.215	0.367	0.227	0.212	0.213	0.209	0.240	0.262	0.256	0.220	0.302	0.261	0.223	0.220	0.376	0.309
1QQ1 A	C_56	0.303	0.304	0.278	0.262	0.215	0.245	0.236	0.277	0.262	0.214	0.228	0.313	0.263	0.217	0.236	0.352	0.350
1C4A A	C_17	0.330	0.338	0.236	0.237	0.245	0.230	0.237	0.306	0.265	0.261	0.299	0.403	0.445	0.308	0.222	0.273	0.336
1NGS A	C_40	0.332	0.289	0.247	0.220	0.253	0.228	0.253	0.229	0.229	0.267	0.266	0.323	0.313	0.261	0.238	0.362	0.527
1NGS B	C_41	0.332	0.289	0.247	0.220	0.253	0.228	0.253	0.229	0.229	0.267	0.266	0.323	0.313	0.261	0.238	0.362	0.527
1QRK B	C_57	0.360	0.459	0.230	0.233	0.216	0.218	0.187	0.286	0.271	0.350	0.263	0.328	0.318	0.207	0.218	0.386	0.483
1EGU A	C_27	0.517	0.535	0.553	0.390	0.371	0.398	0.391	0.374	0.453	0.571	0.570	0.509	0.548	0.413	0.492	0.478	0.588
1GOF	C_31	0.530	0.534	0.542	0.439	0.409	0.403	0.378	0.454	0.465	0.409	0.422	0.446	0.421	0.405	0.487	0.516	0.509
1QFS A	C_50	0.285	0.375	0.355	0.275	0.241	0.268	0.279	0.271	0.404	0.306	0.317	0.424	0.434	0.251	0.261	0.519	0.532
1EUU	C_28	0.336	0.371	0.367	0.323	0.400	0.355	0.296	0.312	0.298	0.377	0.330	0.325	0.389	0.325	0.324	0.369	0.493
1QME A	C_51	0.436	0.430	0.326	0.312	0.315	0.275	0.294	0.393	0.285	0.432	0.344	0.310	0.444	0.325	0.359	0.523	0.453
1DOT	C_24	0.465	0.458	0.471	0.441	0.461	0.437	0.437	0.423	0.484	0.599	0.556	0.608	0.624	0.487	0.494	0.529	0.590
1E6Y B	C_25	0.408	0.489	0.320	0.346	0.316	0.369	0.330	0.317	0.362	0.332	0.363	0.342	0.347	0.392	0.354	0.437	0.377
1E6Y E	C_26	0.408	0.489	0.320	0.346	0.316	0.369	0.330	0.317	0.362	0.332	0.363	0.342	0.347	0.392	0.354	0.437	0.377
1QF7 A	C_46	0.500	0.422	0.402	0.348	0.325	0.426	0.301	0.321	0.343	0.387	0.377	0.467	0.437	0.288	0.315	0.482	0.416
1QF7 B	C_47	0.500	0.422	0.402	0.348	0.325	0.426	0.301	0.321	0.343	0.387	0.377	0.467	0.437	0.288	0.315	0.482	0.416
1QF7 C	C_48	0.500	0.422	0.402	0.348	0.325	0.426	0.301	0.321	0.343	0.387	0.377	0.467	0.437	0.288	0.315	0.482	0.416
1QF7 D	C_49	0.500	0.422	0.402	0.348	0.325	0.426	0.301	0.321	0.343	0.387	0.377	0.467	0.437	0.288	0.315	0.482	0.416
3DMR	C_60	0.427	0.525	0.390	0.362	0.320	0.397	0.374	0.503	0.342	0.358	0.406	0.411	0.449	0.368	0.548	0.446	0.577
1ADO A	C_9	0.171	0.280	0.168	0.169	0.184	0.183	0.166	0.182	0.175	0.193	0.223	0.206	0.179	0.191	0.167	0.265	0.241
1ADO B	C_10	0.171	0.280	0.168	0.169	0.184	0.183	0.166	0.182	0.175	0.193	0.223	0.206	0.179	0.191	0.167	0.265	0.241

1ADO C	C_11	0.171	0.280	0.168	0.169	0.184	0.183	0.166	0.182	0.175	0.193	0.223	0.206	0.179	0.191	0.167	0.265	0.241
1ADO D	C_12	0.171	0.280	0.168	0.169	0.184	0.183	0.166	0.182	0.175	0.193	0.223	0.206	0.179	0.191	0.167	0.265	0.241
1JDA	C_32	0.301	0.161	0.169	0.202	0.248	0.159	0.166	0.158	0.165	0.150	0.162	0.156	0.176	0.158	0.151	0.225	0.204
1PLU A	C_42	0.235	0.237	0.167	0.201	0.163	0.166	0.166	0.162	0.163	0.163	0.163	0.176	0.186	0.190	0.162	0.203	0.213
1AI4 B	C_13	0.211	0.269	0.344	0.210	0.207	0.185	0.214	0.220	0.183	0.259	0.244	0.281	0.235	0.203	0.225	0.262	0.304
1QBQ B	C_45	0.194	0.270	0.218	0.210	0.206	0.222	0.237	0.231	0.272	0.248	0.259	0.265	0.246	0.210	0.264	0.285	0.299
1DJN A	C_21	0.240	0.347	0.211	0.193	0.273	0.302	0.199	0.226	0.243	0.252	0.285	0.255	0.325	0.227	0.214	0.270	0.433
1DJN B	C_22	0.240	0.347	0.211	0.193	0.273	0.302	0.199	0.226	0.243	0.252	0.285	0.255	0.325	0.227	0.214	0.270	0.433
1DLJ A	C_23	0.236	0.313	0.214	0.176	0.171	0.152	0.167	0.149	0.187	0.205	0.185	0.191	0.229	0.202	0.183	0.275	0.310
1QBI A	C_43	0.279	0.251	0.236	0.162	0.153	0.170	0.186	0.161	0.170	0.184	0.172	0.180	0.170	0.154	0.154	0.242	0.301
1QBI B	C_44	0.279	0.251	0.236	0.162	0.153	0.170	0.186	0.161	0.170	0.184	0.172	0.180	0.170	0.154	0.154	0.242	0.301
1QMG A	C_52	0.199	0.330	0.224	0.146	0.161	0.154	0.167	0.158	0.173	0.163	0.195	0.203	0.180	0.153	0.173	0.267	0.203
1QMG B	C_53	0.199	0.330	0.224	0.146	0.161	0.154	0.167	0.158	0.173	0.163	0.195	0.203	0.180	0.153	0.173	0.267	0.203
1QMG C	C_54	0.199	0.330	0.224	0.146	0.161	0.154	0.167	0.158	0.173	0.163	0.195	0.203	0.180	0.153	0.173	0.267	0.203
1QMG D	C_55	0.199	0.330	0.224	0.146	0.161	0.154	0.167	0.158	0.173	0.163	0.195	0.203	0.180	0.153	0.173	0.267	0.203
1F6W A	C_29	0.199	0.222	0.235	0.156	0.161	0.169	0.197	0.183	0.193	0.171	0.205	0.313	0.208	0.206	0.191	0.211	0.242
1JDF A	C_33	0.212	0.226	0.194	0.172	0.173	0.172	0.173	0.200	0.170	0.196	0.200	0.187	0.193	0.176	0.215	0.196	0.279
1JDF B	C_34	0.212	0.226	0.194	0.172	0.173	0.172	0.173	0.200	0.170	0.196	0.200	0.187	0.193	0.176	0.215	0.196	0.279

Table S6. Reduced values of the reordered data matrix by tw-JCA method for HCC probability

Protein	Name	2DN	4VI	2QD	3ED	2FL	7YF	2EQ	3KR	2HN	2WR	2MI	4ML	2HR	3HQ	2VM	2LI	2VL
1JDF C	C_35	0.212	0.226	0.194	0.172	0.173	0.172	0.173	0.200	0.170	0.196	0.200	0.187	0.193	0.176	0.215	0.196	0.279
1JDF D	C_36	0.212	0.226	0.194	0.172	0.173	0.172	0.173	0.200	0.170	0.196	0.200	0.187	0.193	0.176	0.215	0.196	0.279
1AQ2	C_14	0.228	0.221	0.156	0.146	0.154	0.148	0.176	0.168	0.151	0.202	0.201	0.260	0.184	0.158	0.155	0.232	0.246
1C1D B	C_16	0.193	0.205	0.148	0.153	0.155	0.211	0.152	0.161	0.150	0.148	0.150	0.147	0.201	0.146	0.167	0.224	0.254
1CCW B	C_18	0.172	0.242	0.207	0.151	0.146	0.160	0.165	0.190	0.153	0.251	0.179	0.168	0.219	0.184	0.189	0.182	0.244
1CCW D	C_19	0.172	0.242	0.207	0.151	0.146	0.160	0.165	0.190	0.153	0.251	0.179	0.168	0.219	0.184	0.189	0.182	0.244
1D5T A	C_20	0.278	0.205	0.210	0.175	0.175	0.158	0.158	0.177	0.159	0.164	0.284	0.233	0.203	0.192	0.231	0.237	0.226
1BY7 A	C_15	0.145	0.135	0.141	0.125	0.129	0.126	0.142	0.219	0.150	0.143	0.146	0.286	0.207	0.137	0.155	0.195	0.160
1JRR A	C_38	0.145	0.135	0.141	0.125	0.129	0.126	0.142	0.219	0.150	0.143	0.146	0.286	0.207	0.137	0.155	0.195	0.160
1JET A	C_37	0.222	0.212	0.172	0.132	0.140	0.155	0.147	0.136	0.156	0.164	0.142	0.220	0.145	0.149	0.149	0.196	0.245
3GCB	C_61	0.196	0.198	0.181	0.130	0.140	0.153	0.141	0.156	0.181	0.169	0.173	0.222	0.171	0.144	0.189	0.219	0.272

Table S7. tw-JCA reordered data matrix for the probability of the mutated HBC-related proteins to turn into non-cancer proteins

Protein	Name	2DN	2MI	7YF	2EQ	3KR	3ED	2HN	4ML	2HR	2WR	2QD	2VM	3HQ	2VL	2FL	2LI	4VI
ARHGEF4	C_1	0.397	0.395	0.396	0.219	0.264	0.191	0.369	0.374	0.281	0.350	0.284	0.369	0.291	0.289	0.352	0.273	0.356
FLJ13479	C_4	0.429	0.424	0.431	0.393	0.263	0.389	0.386	0.375	0.066	0.429	0.389	0.357	0.351	0.290	0.341	0.411	0.424
MYOD1	C_8	0.424	0.421	0.385	0.421	0.375	0.393	0.425	0.377	0.414	0.410	0.413	0.411	0.411	0.411	0.403	0.399	0.404
KEAP1	C_6	0.457	0.425	0.427	0.418	0.455	0.406	0.441	0.448	0.454	0.455	0.331	0.354	0.403	0.214	0.404	0.376	0.392
EGFL6	C_3	0.360	0.456	0.440	0.474	0.343	0.316	0.397	0.440	0.449	0.471	0.430	0.468	0.460	0.338	0.402	0.416	0.391
PFC	C_9	0.417	0.421	0.399	0.239	0.314	0.328	0.366	0.400	0.355	0.381	0.377	0.364	0.368	0.361	0.412	0.407	0.414
C22orf19	C_2	0.208	0.205	0.182	0.122	0.127	0.123	0.236	0.110	0.206	0.241	0.138	0.188	0.159	0.059	0.126	0.146	0.150
INHBE	C_5	0.514	0.507	0.519	0.469	0.504	0.530	0.502	0.497	0.468	0.445	0.489	0.507	0.507	0.377	0.447	0.445	0.502
KPNA5	C_7	0.243	0.362	0.355	0.209	0.228	0.308	0.346	0.347	0.384	0.380	0.275	0.320	0.370	0.177	0.225	0.211	0.185
RNU3IP2	C_10	0.256	0.250	0.232	0.145	0.143	0.223	0.267	0.253	0.210	0.217	0.223	0.244	0.164	0.115	0.195	0.178	0.214

Table S8. tw-JCA reordered data matrix for the probability of the mutated HCC-related proteins to turn into non-cancer proteins

Protein	Name	2DN	2WR	2HN	2VM	2MI	2EQ	2FL	4VI	2LI	2VL	3KR	4ML	7YF	2HR	2QD	3HQ	3ED
C6orf29	C_1	0.437	0.475	0.469	0.444	0.401	0.440	0.154	0.202	0.109	0.066	0.451	0.431	0.386	0.469	0.438	0.482	0.392
MKRN3	C_6	0.418	0.429	0.449	0.439	0.468	0.398	0.387	0.391	0.458	0.363	0.403	0.432	0.392	0.450	0.443	0.446	0.384
TBX22	C_9	0.423	0.413	0.414	0.412	0.427	0.346	0.388	0.408	0.378	0.351	0.337	0.355	0.384	0.439	0.378	0.455	0.413
CNTN4	C_2	0.344	0.333	0.343	0.340	0.345	0.352	0.403	0.158	0.230	0.179	0.297	0.375	0.384	0.393	0.378	0.418	0.283
GUCY1A2	C_3	0.240	0.266	0.219	0.202	0.193	0.203	0.119	0.134	0.162	0.120	0.147	0.117	0.233	0.219	0.223	0.218	0.199
K6IRS3	C_4	0.263	0.298	0.300	0.301	0.298	0.142	0.284	0.232	0.231	0.179	0.197	0.259	0.279	0.278	0.273	0.273	0.205
RUNX1T1	C_7	0.271	0.289	0.312	0.294	0.260	0.207	0.251	0.248	0.302	0.190	0.160	0.314	0.312	0.183	0.262	0.243	0.209
SFRS6	C_8	0.324	0.299	0.322	0.316	0.316	0.315	0.305	0.241	0.253	0.269	0.118	0.311	0.292	0.284	0.287	0.322	0.227
LOC157697	C_5	0.268	0.273	0.268	0.279	0.277	0.162	0.218	0.235	0.221	0.233	0.207	0.247	0.220	0.227	0.213	0.230	0.085
UHRF2	C_10	0.193	0.181	0.239	0.256	0.256	0.214	0.157	0.183	0.190	0.143	0.134	0.245	0.239	0.226	0.233	0.260	0.098