BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access

www.bioinformation.net

Hypothesis

Volume 8(8)

Analysis of aminoacids pattern in receptor tyrosine kinase using Boolean Association Rule

Pranjal Kalita[1], Brindha Senthil Kumar[2], Soundararajan Krishnaswamy[3] & Senthil Kumar Nachimuthu[2]*

[1]Department of Information Technology, ICFAI University Mizoram, Aizawl - 796 023, Mizoram, India; [2]Bioinformatics Infrastructure Facility, Department of Biotechnology, Mizoram University, Aizawl - 796 004, Mizoram, India; [3]Center of Excellence in Biotechnology Research, King Saud University, Riyadh, Saudi Arabia; Senthil Kumar Nachimuthu – E-mail: nskmzu@gmail.com; *Corresponding author

Received March 24, 2012; Accepted April 18, 2012; Published April 30, 2012

Abstract:
Cancers are characterized by unrestricted cell division and independency of growth factor and other external signal responsiveness. Eukaryotic parental cells of tumors, on the other hand, constitute tissues and other higher structures like organs and systems and are capable of performing various functions in a highly co-ordinated fashion. Hence, cancer cells may be considered as entities capable of incessant growth and cell division but lacking any evolutionarily advanced intracellular or intercellular regulation. Since receptor tyrosine kinases are highly altered and exist in deregulated/constitutively active forms in cancer cells - achieved through various epigenetic mechanisms - we hypothesize the functional RTKs in cancer cells to resemble their counterparts in more primitive species. Analysis of RTK sequences of various species and of cancer is, therefore, expected to prove this hypothesis. Association rule in data mining can reveal the hidden biological information. This study utilizes the Boolean association rule to mine the occurrence pattern of glycine, arginine and alanine in receptor tyrosine kinases (RTKs) of invertebrates, vertebrates and cancer related vertebrate RTKs based on protein sequence informations. The results reveal that vertebrate cancer RTKs resembles prokaryotes and invertebrate RTKs showing an increasing trend of glycine, alanine and decreasing trend in arginine composition. The aminoacid compositions of vertebrates: invertebrates: prokaryotes: vertebrate cancer with respect to Glycine (>=6.1) were 42.86: 50.0: 85.71: 100%, Alanine (>=6.2) were 10.72: 66.67: 85.71: 100%, whereas Arginine (>=5.9) were 21.43: 16.67: 14.29: 0%, respectively. In conclusion, results from this study supports our hypothesis that cancer cells may resemble lower organisms since functionally cancer cells are unresponsive to external signals and various regulatory mechanisms typically found in higher eukaryotes are largely absent.

Background:
Data mining techniques can be applied to study the behavior of different amino acid in protein sequences. The association rule mining technique is a popularly used data mining technique. Association rule mining involves counting frequent patterns (or associations) in large databases, reporting all that exist above a minimum frequency threshold known as the 'support' **[1].**

The receptor tyrosine kinase (RTK) pathway plays critical roles in growth and division of cells. The RTK family comprises numerous cell-surface receptors that mediate cell growth, differentiation, migration and metabolism **[2].** RTKs have an extracellular portion to which polypeptide ligands bind, a single-pass transmembrane helix, and a cytoplasmic portion containing a protein tyrosine kinase domain that catalyses phosphoryl transfer from ATP to tyrosine (Tyr) residues in protein substrates **[3].** In cancer cells, mutations in the genes encoding RTKs and various epigenetic mechanisms like alternative splicing lead to inappropriate activation of kinases resulting in uncontrolled cell division **[4].**

Amino acid restriction sends normal cells into a quiescent mode, their growth and division cycles being shut down in a reversible manner. Tumour cells usually fail to move out of cycle, the resulting imbalance generally leading to cell death in a matter of days **[5].** Our preliminary studies reveal that the percentage of the amino acids present (except glycine, arginine and alanine), is approximately the same in most of the Receptor Tyrosine Kinase (RTK) protein sequences irrespective of different species or taxa, whether it is vertebrate or invertebrate or cancer sequences. Glycine is a non polar neutral amino acid with hydropathy index -0.4. The amino acid glycine was found to reduce tumour growth in rats. Dietary glycine prevented increases in cell proliferation, a key event in cancer development, suggesting that it may be an effective anti-cancer agent **[6].** Arginine is a nonpolar positively charged amino acid with hydropathy index of -4.5. It is involved in a number of biosynthetic pathways that significantly influence carcinogenesis and tumour biology **[7].** Alanine is a neutral nonpolar amino acid with hydropathy index 1.8. Elevated rates of glucose and alanine turnover and gluconeogenesis from alanine were detected in patients who had advanced lung cancer with weight loss **[8].**

This study attempts to analyse the variations in the occurrence of amino acids glycine, arginine and alanine in RTKs of invertebrates, vertebrates and cancers using association rule mining technique.

Methodology:

Data Source and Data Selection

The complete RTK protein sequences have been collected from NCBI databases (www.ncbi.nlm.nih.gov/) and Swiss Prot. There are 28 vertebrate sequences, 6 invertebrate sequences, 7 prokaryote sequences and 2 cancer sequences. The minimum length vertebrate and invertebrate sequences are 1045 and 799, respectively. Two cancer sequences namely human cancer and mouse cancer are of same length 1620. The ProtParam software from Expasy server (web.expasy.org/protparam/) is used to calculate the protein parameters. The protparam results showed interesting features in glycine, arginine and alanine amino acids, hence these three amino acids have selected as a feature set for the association rule mining.

In this study, we consider each amino acid as an item, the protein sequence as basket that contains items and each taxa or species as one transaction. On these transactions association rule mining technique has been applied to obtain meaningful association among the amino acids, also how frequently the amino acid is present in the transactions. The quantitative value for the items has been mapped to Boolean values, and then Boolean association rule mining techniques has been applied to study the behaviour of the amino acids in the sequences.

Association rule

An 'association rule' is a pair of disjoint item sets. If LHS and RHS denote the two disjoint itemsets, the association rule is written as LHS→RHS i.e LHS and RHS are sets of items, the RHS set being likely to occur whenever the LHS set occurs. The 'support' of the association rule LHS→RHS with respect to a transaction set T is the ratio of support (LHS U RHS)/ T. The 'confidence' of the rule LHS→RHS with respect to a transaction set T is the ratio of support (LHS U RHS)/ support (LHS) **[9].**

Boolean association rule

Boolean values were used to represent the present or absent of the item in transaction. '0' represents the absence of particular item in the transaction and '1' represents presence of particular item in the transaction.

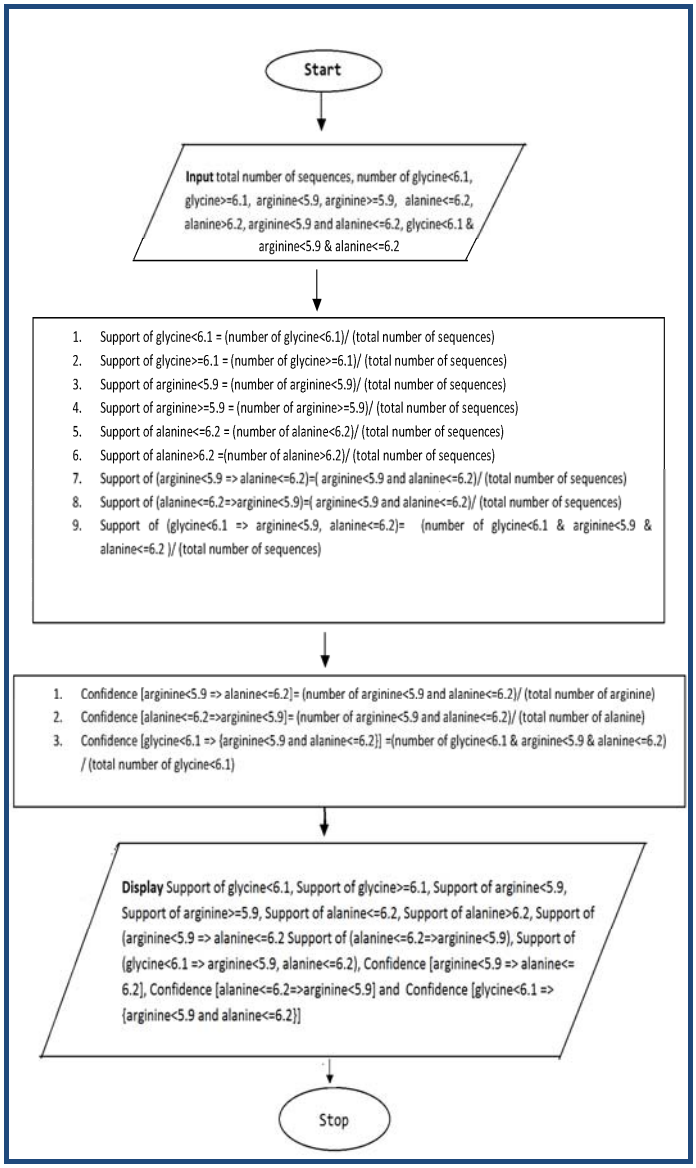

Figure 1: Flow chart of the *in silico* analysis of RTK protein sequences

Data Optimization

The quantitative values of glycine, arginine and alanine columns are converted to boolean form i.e. 0 and 1 **Table 1 (see supplementary material).** Every amino acid column is divided in two groups, the grouping is necessary to convert into Boolean form. Based on the quantitative values, their variations, range etc. the grouping is done like glycine<6.1% and glycine>=6.1%. For example in most of the transactions glycine percent is either very less than 6.1% or more than it. 6.1% can be used as a boundary for this classification. The '1' in the table represents the presence and '0' represent the absence of that item for that particular transaction. Total eight items are been considered (i.e. glycine<6.1%, glycine>=6.1%, arginine<5.9%, arginine>=5.9%, alanine<=6.2% and alanine>6.2%) **(Figure 1).**

Disscussion:

Association rules are used widely in the area of market basket analysis and can also reveal biologically relevant associations between different genes or between environmental effects and gene expression **[9].** The results show that in 42.86% normal vertebrates the glycine composition is more than or equal to 6.1, which is 50% in invertebrates, 85.71% in prokaryotes and 100% in cancer sequences, thereby reflecting the increasing trend of glycine from normal vertebrates to cancerous RTK protein **(Table 1).** Similarly, alanine and arginine show increasing and decreasing trends, respectively, from normal vertebrate sequence to cancer sequences. Correspondingly, the confidence value shows that if the arginine is less 5.9%, then alanine is always less than or equal to 6.2% in vertebrates. Similarly, if glycine is less than 6.1%, then 93.75% alanine is less than or equal to 6.2% and arginine will be less than 5.9%. It describes how one amino acid is associated with another.

Both the human and mouse cancer sequences possess similar characteristics **(Table 1).** It can be seen that the support is either 0 or 100. This reflects that in both the cancer sequence transaction one particular item is either present or absent. The confidence levels show zero for the above-mentioned combinations. It can be assumed that the association between different items is varying from vertebrates to cancer.

RTK activity in resting, normal cells is tightly controlled. When they are mutated or structurally altered, RTKs become potent oncoproteins: abnormal activation of RTKs in transformed cells has been shown to be involved in the development and progression of many human cancers **[10]**. Consequently, RTKs and their growth-factor ligands have become rational targets for therapeutic intervention using humanized antibodies and small molecule drugs. Although a complete understanding of RTK function and dysfunction in diverse tissues and multiple biological processes is still to be achieved, studies of members of this family have already had a significant impact on cancer therapy **[10].**

Analysing the data it is clear that glycine and alanine show increasing trend from normal vertebrate sequences to cancer sequence and on the other hand arginine show the decreasing trend. The association among these three amino acids can be established as follows: support for glycine<6.1 decreases when the support for arginine<5.9 increases. It is also evident that there is a trend in the increase/decrease of amino acid composition from vertebrates to cancer sequences

Conclusion:

In this paper the Boolean association rule mining technique has been applied to find differences in the frequency of incidence of a few important amino acids in various RTKs of different species. The analysis shows that the three amino acid characters – glycine, arginine, alanine - of cancer sequences are more similar towards invertebrates and prokaryotes, which may lead the cancer RTK's to de-evolve.

Acknowledgment:

The authors are thankful to Department of Biotechnology (DBT), Govt. of India, New Delhi for the financial support in the form of Bioinformatics Infrastructure Facility (BTISNeT).

References:

[1] Singh M & Singh G, *Int J Comp Sci Security*. 2011 **5**: 14
[2] Schlessinger J, *Cell*. 2000 **103**: 211 [PMID: 11057895]
[3] Hubbard SR, *Nat Rev Mol Cell Biol*. 2004 **5**: 464 [PMID: 5173825]
[4] Bache KG *et al*. *The EMBO Journal*. 2004 **23**: 2707 [PMID: 15229652]
[5] Wheatley Denys N, *Oncology News*. 2006 **1**: 6 [PMID: 12429007]
[6] Rose ML *et al*. *Carcinogenesis*. 1999 **20**: 793 [PMID: 10334195]
[7] Lind DS, *The Journal of Nutrition*. 2004 **134**: 2837S [PMID: 15465796]
[8] Leij-Halfwerk S *et al*. *The American J Clinical Nutrition*. 2000 **71**: 583 [PMID: 6345246.001]
[9] Anandhavalli M *et al*. *Int J Comp Theory Eng*. 2010 **2**: 269
[10] Gschwind A *et al*. *Nature Rev Cancer*. 2004 **4**: 361 [PMID: 15122207]

Edited by P Kangueane
Citation: **Kalita *et al.*** Bioinformation 8(8): 344-347 (2012)

BIOINFORMATION

open access

Supplementary material:

Table 1: Amino acid transactions in RTK sequences

Transaction ID- Vertebrates	**Glycine: <6.1%**	**>=6.1%**	**Arginine: <5.9%**	**=5.9%**	**Alanine: <=6.2%**	**>6.2%**
Ma's night monkey	1	0	1	0	1	0
Black-handed spider monkey	1	0	1	0	1	0
Red-bellied titi	1	0	1	0	1	0
white-tufted-ear marmoset	1	0	1	0	1	0
Dog	0	1	0	1	0	1
Vervet Monkey	1	0	1	0	1	0
Guereza	1	0	1	0	1	0
common carp	0	1	1	0	1	0
Zebrafish	1	0	1	0	1	0
Horse	1	0	1	0	1	0
Black Lemur	1	0	1	0	1	0
domestic cat	1	0	1	0	1	0
Chicken	0	1	1	0	1	0
Western lowland gorilla	1	0	1	0	1	0
Human	0	1	0	1	0	1
golden hamster	0	1	0	1	0	1
gray mouse lemur	0	1	1	0	1	0
house mouse	0	1	0	1	1	0
Northern white-cheeked gibbon	0	1	1	0	1	0
Rabbit	0	1	1	0	1	0
Sheep	0	1	1	0	1	0
olive baboon	1	0	1	0	1	0
Sumatran orang-utan	1	0	1	0	1	0
Norway rat	0	1	0	1	1	0
Bolivian squirrel monkey	1	0	1	0	1	0
Pig	1	0	1	0	1	0
green pufferfish	1	0	0	1	1	0
Swordtail platyfish	0	1	1	0	1	0

Support (Glycine<6.1%) =57.14%; (Glycine>=6.1%) =42.86% *Support (Alanine<=6.2%) =89.28%; (Alanine>6.2%) =10.72%*

Support (Arginine<5.9%) =78.57%; (Arginine>=5.9%) =21.43%

Arginine<5.9 => Alanine<=6.2 [Support = 78.57% confidence =100%] *Glycine<6.2 => {Arginine<5.9 ,Alanine<=6.2} [Support=53.57% confidence=93.75%]*

Alanine<=6.2 => Arginine<5.9 [Support = 78.57% confidence =88%]

Transaction - Invertebrates	**Glycine: <6.1%**	**>=6.1%**	**Arginine: <5.9%**	**>=5.9%**	**Alanine: <=6.2%**	**>6.2%**
yellow fever mosquito	0	1	1	0	1	0
honey bee	0	1	1	0	0	1
domestic silkworm	1	0	1	0	1	0
southern house mosquito	1	0	1	0	0	1
fruit fly	1	0	1	0	0	1
black-legged tick	0	1	0	1	0	1

Support (Glycine<6.1%) =50%; (Glycine>=6.1%) =50% *Support (Alanine<=6.2%) =33.33%; (Alanine >6.2%) =66.67%*

Support (Arginine<5.9%) =83.33%; (Arginine>=5.9%) =16.67%

Arginine<5.9 => Alanine<=6.2 [Support = 33.33% confidence =40%] *Glycine<6.2 => {Arginine<5.9 ,Alanine<=6.2} [Support=16.67% confidence=33.33%]*

Alanine<=6.2 => Arginine<5.9 [Support = 33.33% confidence =100%]

Transaction- Prokaryotes	**Glycine: <6.1%**	**>=6.1%**	**Arginine:<5.9%**	**>=5.9%**	**Alanine:<=6.2%**	**>6.2%**
Lactobacillus johnsonii	0	1	1	0	1	0
Escherichia coli	0	1	1	0	0	1
Pseudomonas syringae pv. P	1	0	0	1	0	1
Salmonella typhimurium	0	1	1	0	0	1
Streptococcus suis	0	1	1	0	0	1
Streptococcus suis	0	1	1	0	0	1
Streptococcus uberis	0	1	1	0	0	1

Support (Glycine<6.1%) =14.29%; (Glycine>=6.1%) =85.71% *Support (Alanine<=6.2%) =14.29%; (Alanine>6.2%) =85.71%*

Support (Arginine<5.9%) =85.71%; (Arginine>=5.9%) =14.29%

Arginine<5.9 => Alanine<=6.2 [Support = 33.33% confidence =40%] *Glycine<6.2 => {Arginine<5.9 ,Alanine<=6.2} [Support=16.67% confidence=33.33%]*

Alanine<=6.2 => Arginine<5.9 [Support = 33.33% confidence =100%]

Transaction – Vertebrate Cancer sequences	**Glycine: <6.1%**	**>=6.1%**	**Arginine:<5.9%**	**>=5.9%**	**Alanine: <=6.2%**	**>6.2%**
Human Cancer	0	1	1	0	0	1
Mouse Cancer	0	1	1	0	0	1

Support (Glycine<6.1%) =0%; (Glycine>=6.1%) =100% Support (Alanine<=6.2%) =0%; *(Alanine>6.2%) =100%*

Support (Arginine<5.9%) =100%; (Arginine>=5.9%) =0%

Arginine<5.9 => Alanine<=6.2 [Support = 0% confidence =0%] *Glycine<6.2 => {Arginine<5.9 ,Alanine<=6.2} [Support=0% confidence=0%]*

Alanine<=6.2 => Arginine<5.9 [Support = 0% confidence =0%]