JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE

In silico directed mutagenesis using software for glycosylation sites prediction as a new step in antigen design

Vladislav Victorovich Khrustalev*[1] and Eugene Victorovich Barkovsky[1].

[1]Department of General Chemistry, Belarussian State Medical University, Belarus, Minsk, Dzerzinskogo, 83

Received: 29 September 2011 **Accepted:** 09 February 2012 **Available Online**: 22 February 2012

ABSTRACT

In silico directed mutagenesis with the aim to estimate consequences of mutational pressure on number of N- and O-glycosylation sites has been proposed as an important step in antigen design. Using this kind of methodology one is able to estimate probabilities at which N- and O -glycosylation sites can be destroyed and created due to one-step missense mutations caused by mutational pressure in the subsequent gene and to select a region of a protein with a lowest probability of new glycosylation site occurrence. Mutational AT-pressure has been simulated in the region of *env* gene coding for HIV1 gp120. Consequences of 741 amino acid substitutions caused by missense GC to AT mutations have been predicted with the help of NetNGlyc 1.0 and NetOGlyc 3.1 algorithms. The probability of O-glycosylation site destruction (2.16%) in HIV1 gp120 protein due to a single missense GC to AT mutation in *env* gene is higher than the probability of a new site creation (0.40%). The probability of N-glycosylation site destruction in HIV1 gp120 protein is equal to the probability of its creation (5.53%), while the number of N-glycosylation sites which can be created due to a single missense GC to AT mutation in *env* gene is 1.27 times higher than the number of N-glycosylation sites which can be destroyed

Keywords: HIV1; N-glycosylation; O-glycosylation; gp120; mutational pressure; B-cell epitopes .

1. Introduction

Practical approach of the mutational pressure theory created by Noboru Sueoka [1] has been developed in our recent work [2]. Levels of mutability for different conformational B-cell epitopes of the same protein can be compared based on knowledge of mutational pressure directions in the subsequent gene [2]. The less mutable epitope is encoded by a part of a gene which is protected from missense nucleotide mutations occurrence better than others due to certain features of its nucleotide content and composition [2].

According to our results, there is AT-pressure in the *env* gene of Human immunodeficiency virus type 1 (HIV1), while the rates of G to A transitions in that gene are higher than rates of C to T(U) transitions, and the rates of C to A transversions are higher than rates of G to T(U) transversions [2].

Symmetric mutational AT-pressure leads to the decrease of the quantity and length of linear B-cell epitopes [3, 4]. However, antigenic properties of glycoproteins depend both on existence of immunogenic amino acid stretches forming conformational epitopes and on existence of glycans connected with them [5]. Glycan can be connected with protein via "-OH" group of serine or threonine side chain (this type of glycosylation is known as O-glycosylation) and via "-NH_2" group of asparagine side chain (this type of glycosylation is known as N-glycosylation) [5, 6]. Side chains of some other natural and modified amino acids can also be glycosylated. It is known that many hydroxylisine residues of collagen are O-glycosylated [5]. Rare cases of O-glycosylation via tyrosine and hydroxyproline side chains

***Corresponding author:** Vladislav Victorovich Khrustalev; address: Belarus, Minsk, 220029, Communisticheskaya 7-24; telephone: 80172845957; E-mail Address: vvkhrustalev@mail.ru

have been described, as well as a single (to this date) case of N-glycosylation via arginine side chain [5, 6]. Software for prediction of sites for those rare types of O- and N-glycosylation has not been created yet due to the insufficient volume of data.

The purpose of this study was to simulate AT-pressure in the region of *env* gene coding for HIV1 gp120, to estimate its influence on frequencies of N-glycosylation (via asparagine residues) and O-glycosylation (via serine and threonine residues) sites and to select B-cell epitope of gp120 with lowest probability of a new glycosylation site occurence.

Sites for O-glycosylation have been predicted by NetOGlyc 3.1 software [7] (http://www.cbs.dtu.dk/services/NetOGlyc). This software producing neural network predictions based on 299 known and verified mucin-type O-glycosylation sites has already been used in theoretical study [8].Sites for N-glycosylation have been predicted by NetNGlyc 1.0 software (http://www.cbs.dtu.dk/services/NetNGlyc), which has also showed good performance in several bioinformatical works [9, 10, 11].

Peptides corresponding to those B-cell epitopes of viral or bacterial proteins which are not glycosylated and have low probability to acquire a new site for glycosylation due to mutational pressure should show better performance as antigens for ELISA test systems and as components of synthetic vaccines. Immunization to that kind of peptides should decrease the probability of immune escaping by the way of mutation creating a new glycosylation site. Those B-cell epitopes which can acquire N- or O-glycosylation site at high probability due to a single amino acid substitution caused by mutational pressure should be excluded from antigen design study on its early *in silico* step. It will help future investigators to save time and funds for their *in vitro* and *in vivo* experiments (peptide synthesis, testing antigenic properties of peptides in ELISA, immunization of laboratory animals, affinity purification of antibodies and, finally, clinical trials).

2. Material and Methods

As a material we used nucleotide sequence of the HIV1 *env* gene region coding for gp120 protein from the reference strain of that virus [NC_001802]. We introduced all possible point missense nucleotide mutations of GC to AT direction in that coding region. Then consequences of each of those possible 741 amino acid substitutions have been predicted with the help of NetOGlyc 3.1 [7] and NetNGlyc 1.0 algorithms.

Introduction of point missense nucleotide mutations of GC to AT direction has been performed with a help of simple but useful original MS Excel algorithm entitled "Mutational Pressure Simulator". To use this software available via our webpage (www.barkovsky.hotmail.ru) one should enter nucleotide sequence of a protein coding region in a special cell on its "nucleotide sequence" list. Then one should enter certain codon in which mutation should be introduced as well as resulting codon in cells on the same list. The set of amino acid sequences with introduced mutations can be found in a column of the "nucleotide sequence results" list. Each of those sequences possesses amino acid substitution resulting from single codon mutation. The algorithm uses universal genetic code to translate nucleotide sequences into amino acid sequences. In case if there are some deviations from universal genetic code in the genome of given specie, the code used by the algorithm may be changed manually. To make this operation one should introduce corrections into the genetic code table on the "genetic code" list. The output of the "Mutational Pressure Simulator" algorithm (amino acid sequences) is in FASTA format. It means that all the resulting sequences may be copied from the "nucleotide sequence results" list and pasted into the special field of NetNGlyc 1.0 or NetOGlyc 3.1.

"Mutational Pressure Simulator" is also able to introduce single amino acid mutations in the amino acid sequence entered in the cell on the "amino acid sequence" list. The set of sequences with introduced mutations can be found in a column of the "AA results" list in FASTA format.

Information from the output of NetNGlyc 1.0 and NetOGlyc 3.1 used in this study includes: 1) number of sites for N- and O-glycosylation in each mutated sequence; 2) location of a new N-glycosylation or O-glycosylation site; 3) location of a site for N-glycosylation or O-glycosylation which has been destroyed due to a single amino acid substitution.

To calculate total probability of N-glycosylation site creation due to a single amino acid substitution caused by AT-pressure we divided the number of amino acid substitutions creating new sites for N-glycosylation by the total number of possible amino acid substitutions produced by missense GC to AT mutations. Other probabilities, including those for each glycosylation site destruction and creation, have been calculated in a similar way.

3. Results

3.1 *Consequences of GC to AT missense mutations in env gene for number of O-glycosylation sites in HIV1 gp120 protein*

According to the results of NetOGlyc 3.1 prediction, there is a single site for O-glycosylation in HIV1 gp120 protein from reference strain. This site is quite unstable under the pressure of GC to AT nucleotide mutations (see Figure 1). Sixteen amino acid substitutions in the area near that site lead to its disappearance (the probability of its destruction due to a single missense GC to AT mutation is equal to 2.16%). Just two of those mutations lead to the replacement of threonine residue itself (See Supplementary material, Tables 1-4). Six of those mutations lead to the replacement of alanine residue near threonine, and four – to the replacement of proline residue. This data is in consistence with the known fact that threonine and serine residues are O

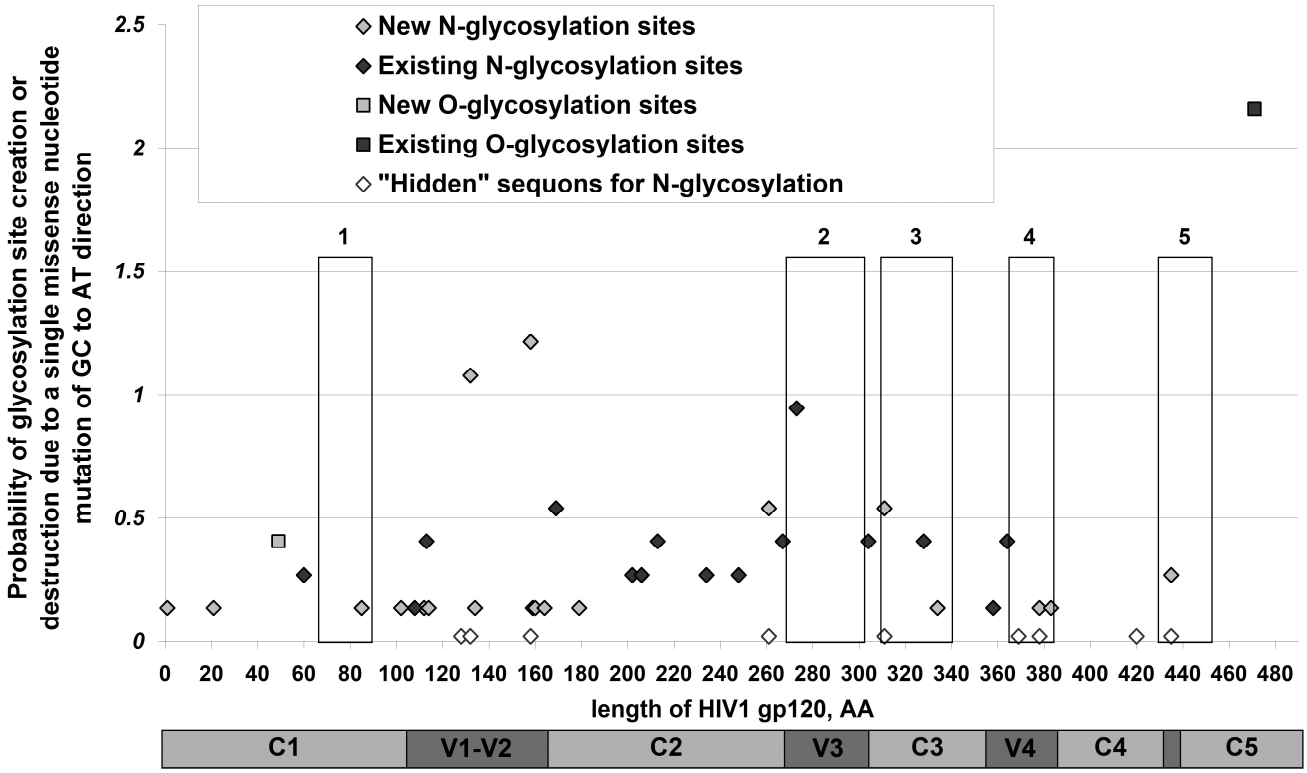

Figure 1. Probabilities of creation and destruction for N- and O-glycosylation sites along the length of gp120 protein from HIV1 reference strain. Five conformational epitopes [2] are designated by bars. Boarders of canonical conserved (C) and variable (V) regions of gp120 are provided. "Hidden sequons" for N-glycosylation are shown

-glycosylated in case if they are surrounded by proline and alanine residues [7]. Both proline and alanine are encoded by GC-rich codons (CCX and GCX, respectively).

Interestingly, only a single O-glycosylation site can be created in gp120 protein due to one-step missense GC to AT mutations in the region of *env* gene coding for it. The probability of creation for that O-glycosylation site (0.40%) is 5.4 times lower than the probability of destruction for existing O-glycosylation site. New O-glycosylation site can appear in the sequence "P**T**DPNP" enriched by proline residues situated in C1 region of gp120.

3.2 *Consequences of GC to AT missense mutations in env gene for number of N-glycosylation sites in HIV1 gp120 protein*

According to the results of NetNGlyc 1.0 predictions, the probability of N-glycosylation site creation in HIV1 gp120 protein due to a single missense GC to AT mutation in *env* gene (5.53%) is exactly the same as the probability of N-glycosylation site destruction. Since asparagine is encoded by GC-poor codons (AAT/C), its level of usage increases due to GC to AT mutations. However, consensus sequence for N-glycosylation site (also known as "sequon") is "Asn-Xaa-Thr/Ser" (where Xaa is not Pro). It means that serine and threonine replacements due to AT-pressure lead to the destruction of N-glycosylation sequons.

There are fifteen sites for N-glycosylation in gp120 protein (according to NetNGlyc 1.0 prediction), each of which can be destroyed by AT-pressure due to replacement of serine or threonine residue in its sequon. However, some of those sites can be destroyed due to many other amino acid replacements (See Supplementary material, Tables 1-4). The most unstable N-glycosylation site is situated in the V3-loop of gp120: it can be destroyed due to seven different amino acid replacements caused by AT-pressure both in the sequon itself and in the area near it (see Figure 1).

The number of N-glycosylation sites which can appear due to AT-pressure is 1.27 times higher than the number of sites which can be destroyed. Indeed, since serine and threonine are encoded by codons of average GC-content, they can disappear and also appear due to GC to AT missense mutations. So, in general, AT-pressure leads to the increase of the number of N-glycosylation sites in gp120 protein.

3.3 *Consequences of different types of GC to AT missense mutations in env gene for number of N-glycosylation sites in HIV1 gp120 protein*

For 215 possible missense G to A transitions (see Supplementary material, Table 1) the probability of N-glycosylation site creation is higher than that for N-glycosylation site destruction (4.19% versus 2.33%). Those one-step transitions can create nine and destroy four N-glycosylatyion sites. Since G to A transitions occur in HIV1 genes more frequently than other types of GC to AT mutations, in general, this specific mutational A-pressure [12] should lead to the growth of the number of N-glycosylation sites in gp120 protein. Obviously, amino acid substitutions of Ser2 to Asn direction can not only create new sites for N-glycosylation due to Asn appearance, but also destroy previously existed ones due to Ser disappearance (See Supplementary material, Table 1).

For 128 possible missense C to T(U) transitions (see Supplementary material, Table 2) the probability of N-glycosylation site creation is lower than that for N-glycosylation site destruction (5.47% versus 11.72%). Those one-step transitions can create four and destroy ten N-glycosylatyion sites. Two thirds of destructive missence C to T(U) mutations lead to Thr to Ile replacement (see Supplementary material, Table 2). Sequons for N-glycosylation containing threonine in their third positions can be destroyed by this kind of amino acid substitution, as well as by Thr to Met substitution which may also be caused by missence C to T(U) mutation.

For 187 possible missense C to A transversions (see Supplementary material, Table 3) the probability of N-glycosylation site creation is also lower than that for N-glycosylation site destruction (7.49% versus 8.56%). Those one-step transversions can create nine and destroy twelve N-glycosylatyion sites. The main cause of the destructive effect of missence C to A transversions is in their ability to cause Thr to Lys amino acid substitutions which can destroy sequons for N-glycosylation (see Supplementary material, Table 3).

For 220 possible missense G to T(U) transversions (see Supplementary material, Table 4) the probability of N-glycosylation site creation is higher than that for N-glycosylation site destruction (5.00% versus 2.28%) mostly due to their ability to cause Lys to Asn substitutions. Those one-step transversions can create eight and destroy four N-glycosylatyion sites.

The probability to be N-glycosylated for a given site may become low not only due to direct destruction of the sequon (due to Asn, Ser or Thr disappearance), but also due to other types of amino acid substitutions. Theoretically, some of those amino acid substitutions should be able to make the region containing sequon less hydrophilic. For example, Pro to Leu4 substitution caused by C to T(U) mutation drastically decreased the score for N-glycosylation of a certain sequon (see Supplementary material, Table 2). As we have found out, C to T(U) mutations are prone to decrease the score of linear B-cell epitopes predicted by BepiPred 1.0 [13]. In general, missense C to T(U) mutations should have a destructive effect on antigenic determinants (both glycosylated and not glycosylated) helping the virus to escape humoral immune answer [3]. However, the percent of destructive amino acid replacements which are unable to destroy a sequon for N-glycosylation is relatively low (7.3%). In contrast, the percent of amino acid substitutions which are unable to create a new sequon but have made the score of the sequon higher than the threshold is equal to 53.7%. In other words, mutational AT-pressure destroys sequons for N -glycosylation mostly in direct manner (causing replacements of Asn, Ser and Thr), while it can 1) create new sequons for N-glycosylation and 2) make “hidden sequons” more suitable for that kind of posttranslational modification. “Hidden sequons” themselves may also be consequences of AT-pressure which increases the level of asparagine usage in proteins. They should be abundant in proteins encoded by GC-poor genes, such as HIV1 gp120 protein. Indeed, there are 24 sequons for N-glycosylation in the gp120 protein. All of those sequons may be glycosylated *in vitro* [14]. However, according to NetNGlyc 1.0 predictions, only 15 of them have a high probability to be glycosylated.

4. Discussion

4.1 *The role of in silico directed mutagenesis using software for glycosylation sites prediction in antigen design studies*

The final aim of antigen design study is in the development of new components for vaccines. Synthetic and recombinant vaccines are usually based on peptides corresponding to short fragments of proteins exposed on a surface of virions or bacterial cells. The main idea of synthetic vaccine creation is in the possibility to immunize person against the conserved antigenic determinant shared by the most of the strains of the given pathogen. To make this idea work well one should carefully select antigenic determinant of the protein of interest. The first step of that selection usually includes mapping of B-cell epitopes. There are numerous bioinformatical methods able to predict B-cell epitopes (either liner [13, 15, 16] or conformational [17, 18]) or regions of a protein exposed to a solvent [19]. However synthetic peptides corresponding to strongest B-cell epitopes often can not be recognized by antibodies against fragments of the native molecule and *vice versa*. One of the causes of the lack of immunological cross-reactivity is in the fact that many strong B-cell epitopes are mapped in regions containing proline and glycine residues [4]. Those residues usually form beta-turns which are situated on a surface of a molecule. However those beta-turns are formed mostly during the folding of the whole molecule. Short peptides corresponding to them usually have quite different conformations. So, one should select those fragments of a protein which conformation and secondary structure should not depend on long-distance interactions with other parts of a native full-length molecule, even though their score of immunogenicity is some lower than for other fragments [2].

The second step of antigen design study usually includes

searching for conserved B-cell epitopes in a protein of interest. This step is usually based on alignment of amino acid sequences of that protein from different strains of the same pathogen [2]. As a result, B-cell epitope the structure of which is under the influence of the stronger negative selection should be determined.

In our works [2, 20] we introduced another step of antigen design studies based on mutational pressure theory. The aim of this step is in the selection of the less mutable B-cell epitope. To select the less mutable B-cell epitope one should estimate the most frequent types of nucleotide mutations in a gene coding for a protein of interest (main directions of the mutational pressure) and compare levels of mutability for regions coding for B-cell epitopes. The less mutable B-cell epitope should be encoded by a region of a gene with the lowest level of missense sites for the most frequent nucleotide mutations and the highest level of synonymous sites for them [2]. The probability to be missense for the most frequent types of nucleotide mutations should be low in a region of a gene coding for the less mutable B-cell epitope [2].

In the current article we introduce yet another important step of antigen design study which is also based on mutational pressure theory [1]. It is known that appearance of new sites for glycosylation in viral epitopes leads to the loss of protective effect in case of immunization against recombinant or synthetic peptides [21]. That is why it is important to estimate a risk of a new site for glycosylation appearance due to mutations of a preferable direction in each of the conformational epitopes of a protein of interest before the start of the vaccine design project. Synthetic peptides corresponding to B-cell epitopes which are glycosylated or have a high probability to be glycosylated after a single amino acid substitution caused by mutational pressure should not be used as antigens for vaccine development.

In case if all the conformational epitopes already possess sites for glycosylation, an epitope with the most stable site for glycosylation should be chosen for production of recombinant peptide in eukaryotic cells. Synthetic peptide can also be conjugated with a certain glycan. This kind of synthetic vaccine may be developed in case if the structure of a glycan is constant. It has been shown that different glycans are attached to different N-glycosylation sites of HIV1 gp120 protein in different cell lines [21, 22].

4.2 *Selection of the best antigen for vaccine design study using results of in silico directed mutagenesis with software for glycosylation sites prediction*

Five conformational B-cell epitopes have been mapped by us on HIV1 gp120 protein in the previous study [20]. For this purpose DiscoTope 1.2 [17], Epces [18] and Epitopia [19] algorithms have been used. In the current work we compared their suitability for inclusion in antigen design study based on data received from *in silico* directed mutagenesis session with software for glycosylation sites prediction.

In Figure 1 probabilities of destruction and creation are given for each N-glycosylation site along the length of gp120 protein from reference HIV1 strain. The highest probability of new N-glycosylation site appearance due to AT-pressure is characteristic to V1-V2 region of gp120.

There are no N-glycosylation sites in two from five conformational B-cell epitopes of gp120 predicted by us [2] (in epitope 1 from C1 region and in epitope 5 from C4-V5-C5 region). AT-pressure can create a single N-glycosylation site in epitope 1 at a probability which is two times lower than that for creation of a single N-glycosylation site in epitope 5. Two N-glycosylation sites can appear due to AT-pressure in epitope 3 (this region plays important role in CD4 receptor binding) and epitope 4 (in highly variable V4-loop).

Five sites for N-glycosylation which can appear due to several single amino acid substitutions are represented by "hidden sequons". Different types of amino acid substitutions near those sequons may drastically increase their probabilities to be N-glycosylated. In contrast, probabilities to be N-glycosylated for three other "hidden sequons" did not become higher than the threshold during the current session of *in silico* mutagenesis.

"Hidden sequons" which may become N-glycosylated due to a single amino acid substitution caused by mutational AT-pressure can be found in epitope 3, epitope 4 and epitope 5 (see Figure 1). Moreover, epitope 4 possesses yet another "hidden sequon" which cannot become suitable for N-glycosylation at least due to a single amino acid substitution caused by mutational AT-pressure.

Results showed that epitope 1 is more suitable for antigen design than four other epitopes predicted by us [2]. That region is not glycosylated, it has no "hidden sequons" for N-glycosylation and the probability of N-glycosylation site creation due to AT-pressure is lower for it than for other conformational B-cell epitopes. Moreover, this epitope is the less mutable one [2].

Results of our *in vitro* experiments showed that there are antibodies able to cross-react with the peptide NQ21 corresponding to the consensus sequence of that epitope in blood of 80.22% of persons with currently diagnosed HIV1-infection [2]. This high level of sensitivity for an ELISA test system based on the short NQ21 peptide conjugated with biotin could not be reached in case of a high probability of new N-glycosylation site appearance in the epitope 1 of gp120.

5. Concluding Remarks

In silico directed mutagenesis with estimation of the mutational pressure consequences on number of N- and O-glycosylation sites is an important step in the process of antigen design. Short peptides corresponding to those epitopes which are not glycosylated and have a lowest

probability of new N-glycosylation site appearance due to mutational pressure are recommended for usage as new vaccine components and antigens for ELISA test systems.

6. Supplementary material

Supplementary data and information is available at: http://www.jiomics.com/index.php/jio/rt/suppFiles/80/0
Table 1.Consequences of G to A missense mutations.
Table 2. Consequences of C to T(U) missense mutations.
Table 3. Consequences of C to A missense mutations.
Table 4. Consequences of G to T(U) missense mutations.

References

1. N. Sueoka, Proc. Natl. Acad. Sci. USA 85 (1988) 2653–2657.
2. V. V. Khrustalev, E.V. Barkovsky, A.E. Vasilevskaya, S.M. Skripko, V.L. Kolodkina, G.M. Ignatyev, P.A. Semizon, JIOMICS 1 (2011) DOI: 10.5584/jiomics.v2011i2011.64
3. V. V. Khrustalev, Molecular Immunology 47 (2010) 1635-1639. doi:10.1016/j.molimm.2010.01.006
4. V. V. Khrustalev, E.V. Barkovsky, J. Theor. Biol., 282 (2011) 71-79. doi:10.1016/j.jtbi.2011.05.018
5. K. Drickamer, M.E. Taylor, Introduction to Glycobiology, second ed., Oxford University Press, USA, 2006.
6. M. Lommel, S. Strahl, Glycobiology, 19 (2009) 816–828.
7. K. Julenius, A. Mølgaard, R. Gupta and S. Brunak, Glycobiology, 15 (2005) 153-164.
8. J. J Calvete, L. Sanz, Methods Mol. Biol. 446 (2008) 281-292.
9. L. Wang, F. Li, W. Sun, S. Wu, X. Wang, L. Zhang, D. Zheng, J. Wang and Y. Gao, Mol. and Cell. Proteomics 5 (2006) 560-562. doi: 10.1074/mcp.D500013-MCP200
10. S. K. Saxena, N. Mishra, R. Saxena, M. L. A. Swamy, P. Sahgal, S. Saxena, S. Tiwari, A. Mathur, M. P. Nair, J. Infect. Dev. Ctries 4 (2010) 1-6.
11. A. K. Tomar, B. S. Sooch, S. Yadav, Bioinformation 7 (2011) 69-75.
12. B. Berkhout, F.J. van Hemert, Nucleic Acids Res., 22 (1994) 1705–1711.
13. J. E. P. Larsen, O. Lund, M. Nielsen, Immunome Res. 2 (2006) 2.
14. C. K. Leonard, M. W. Spellman, L. Riddle, R. J. Harris, J. N. Thomas and T. J. Gregory, J. Biol. Chem. 265 (1990) 10373-10382.
15. T. P. Hopp, K. R. Woods, Mol. Immunol. 20 (1983) 483–489.
16. J. Kyte, R. Doolittle, J. Mol. Biol. 157 (1982) 105-132.
17. P. H. Andersen, M. Nielsen, O. Lund, Protein Science 15 (2006) 2558–2567.
18. S. Liang, D. Zheng, C. Zhang, M. Zacharias, BMC Bioinformatics 10 (2009) 302.
19. N. D. Rubinstein, I. Mayrose, E. Martz, T. Pupko, BMC Bioinformatics 10 (2009) 287.
20. V. V. Khrustalev, Immunological Investigations 39 (2010)