Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W99–W104***
doi:10.1093/nar/gki421

WebProAnalyst: an interactive tool for analysis of quantitative structure–activity relationships in protein families

Vladimir A. Ivanisenko[1,2,*], Alexey M. Eroshkin[3] and Nickolay A. Kolchanov[1]

[1]Institute of Cytology and Genetics SBRAS, Lavrentyev avenue 10, Novosibirsk, 630090, Russia, [2]The Novosibirsk State University, Novosibirsk, 630090, Russia and [3]State Research Center of Virology and Biotechnology VECTOR, Koltsovo, Novosibirsk Region, 630559, Russia

Received February 14, 2005; Revised and Accepted March 22, 2005

ABSTRACT

WebProAnalyst is a web-accessible analysis tool (http://wwwmgs.bionet.nsc.ru/mgs/programs/panalyst/) designed for scanning quantitative structure–activity relationships in protein families. The tool allows users to search correlations between protein activity and physicochemical characteristics (i.e. hydrophobicity or alpha-helical amphipathicity) in queried sequences. WebProAnalyst uses aligned amino acid sequences and data on protein activity (pK, K_m, ED_{50}, among others). WebProAnalyst implements methods of the known ProAnalyst package, including the multiple linear regression analysis and the sequence–activity correlation coefficient. In addition, WebProAnalyst incorporates a method based on neural networks. The WebProAnalyst reports a list of sites in protein family, the regression analysis parameters (including correlation values) for the relationships between the amino acid physicochemical characteristics in the site and the protein activity values. WebProAnalyst is useful in search of the amino acid residues that are important for protein function/activity. Furthermore, WebProAnalyst may be helpful in designing the protein-engineering experiments.

INTRODUCTION

When doing studies in functional genomics, there looms the problem of how the molecular structure of a protein relates to its biological effects (1). In fact, the presence of a small number of functionally important residues is the hallmark feature of the biological activity of a protein. The physicochemical, evolutionary and/or structural characteristics of the residues may suffice to describe the variations in the activities of proteins in homologous groups.

One way to predict functionally important residues is the identification of conserved residues in protein groups and their related functional specificity and/or evolutionary trace (2,3). The idea is to identify the columns in the multiple alignment in which the amino acid distribution is closely associated with the grouping by protein specificity. The SDPpred program has proved to be useful in this respect by ensuring the identification of residues that account for protein specificity (4).

Another way to rank protein/peptide residues by functional importance relies on the quantitative analysis of the structure/sequence–activity relationships. For this purpose, the statistical models (multiple linear regression analysis, neural networks, projections to latent structures) are advantageous because they relate protein activity to variables that describe protein site properties, e.g. alpha-helicity, hydrophobicity/hydrophilicity, charge, among others (5–7).

The ProAnalyst methods have been elaborated for the analysis of quantitative data on protein activities (8–10). The ProAnalyst program provides automated generation and verification of the hypotheses on quantitative relationships between the physicochemical characteristics in the regions of aligned protein sequences and their activities. ProAnalyst is multipurpose: it queries for a region, whose substitutions are correlated with variations in the activities of a set of homologous proteins, the so-called activity-modulating sites; it searches for the key physicochemical characteristics that affect the changes in the activities; and it enables the building of multiple linear regression models that relate these characteristics to protein activities. ProAnalyst is provided with a DOS/Windows interface. WebProAnalyst incorporates the major ProAnalyst methods and ensures their access through a web interface. Users can also have recourse to the neural networks method. An advantage of WebProAnalyst is that it enables users to interpret the differences in protein activities in the broad terms of physicochemical properties of the

*To whom correspondence should be addressed. Tel: +7 3832332971; Fax: +7 3832331278; Email: salix@bionet.nsc.ru

activity-modulating centres, to predict protein activities and also to redesign proteins with addressed modified activities.

We queried an exemplary protein family and a set of mutant peptides to demonstrate how WebProAnalyst works. The issue of how the genotype (the structure of M2 proteins) and the phenotype (drug resistance and susceptibility) of a virus may be related was considered for influenza A viruses. A linear dependence between antimicrobial activity and peptide amphipathicity was established for a set of mutant histatin analogues.

MATERIALS AND METHODS

Sequence–activity correlation coefficient

We developed the sequence–activity correlation/determination coefficient (SACC/SADC) as a guide in our search for the functionally important positions in a multiple alignment of homologous proteins. The method has been described in detail preivously (8). Briefly, the SADC may be defined as the proportion of the variation in the protein activities explicable by amino acid substitutions at positions of a multiple alignment. The SACC is calculated as the square root of the SADC and expresses the strongest multiple correlation between the physicochemical characteristics of a site in a multiple alignment and the protein activities. We proceeded on the statistical procedure of regression analysis applied to the data of repeated experiments in which independent variables have the same value, while the dependent may assume different values. We set out by applying the max R^2, the coefficient introduced by Draper and Smith (11) for calculating the maximally attainable correlation for the repeated experiments. In our case, proteins having matching values of characteristics at a site (the independent variables) are treated as repeated experiments for measuring the protein activities (the dependent variable). Let us now turn to repeated experiments in terms of alphabetical analysis of the amino acid sequences. Let the site be given by an amino acid sequence perceived as a 'word' in an alphabet. The grouping of proteins in such a way that the words in each group are the same yields a set of repeated experiments. All the proteins referred to a group will become the repeated experiments performed for measuring protein activities. Let the proteins be assigned to m groups by the matches between amino acids at a site in a multiple alignment. Thus, the maximum attainable correlation between the protein activities and a set of all the possible combinations of amino acid characteristics of a given site is derived from the max R^2 (8)

$$\mathrm{SACC} = \sqrt{1 - \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{ij}\right)^2}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}\right)^2}},$$

where n_i is the number of proteins of the i-th group, Y_{ij} is the protein activity value of the j-th protein of the i-th group, $\bar{Y}_i$ stands for the mean value of the protein activity of the i-th group and $\bar{Y}$ for the total mean value of the activity for the whole protein set. In fact, if the sites of two proteins have the same amino acid sequences, all the physicochemical characteristics would assume the same values. This means that the variation in activity within a group resulting from protein grouping by site matching cannot be reduced further by additionally involving any set of physicochemical properties of the amino acids at these sites. Thus, the SACC is the upper estimate for the multiple correlation coefficients between the physicochemical characteristic(s) and the protein activity of a site. The SACC/SADC coefficient makes possible the calculation of the value of the possibly highest correlation achievable for the quantitative relationship between the physicochemical properties of the sites and the protein activities. The SACC/SADC is a convenient means for an arrangement of positions by their functional significance. However, the SACC correlation gives neither information nor clue, as to whether the properties are correlated and as to how they may be related to the protein activity. For a detailed annotation of the quantitative structure–activity relationships, WebProAnalyst implements multiple linear regression and neural networks analyses.

Multiple regression and neural networks analyses

The methods of multiple linear regression analysis we apply in WebProAnalyst have been described previously (8–10). WebProAnalyst is now expanded by back-propagation neural networks (12). The current version of this method is implemented as a two-layered network without a hidden layer, one layer as an input and the other as an output. The input data are transformed into the weighted sum that is used as an argument of transfer function giving the ultimate output. The continuous sigmoid transformation function is applied as described previously (12). The input data are the numerical values for the physicochemical characteristics of a site given by a sliding window. The numerical values lie in the 0–1 interval as a result of the standard input transformation. The minimum value of the variable is subtracted from the variable, and the resulting value is divided by the difference between the maximum and the minimum values. The output data are the predicted activity values. The activity values in the course of training of the neural network are also transformed into values that accommodate into the 0–1 range. When testing the neural network, the reported value on the output layer neuron is translated into the activity value by reverse transformation. WebProAnalyst then treats a single activity for a single protein. It should be noted that the accurate predictions require representative training samples.

SOFTWARE ACCESS

The WebProAnalyst web page is available at http://wwwmgs2.bionet.nsc.ru/mgs/programs/panalyst/ (Figure 1). The input WebProAnalyst consists of a multiple sequence alignment, protein activity values, slide window length and physicochemical properties. It is advisable that the length of the sliding window would be of 1–20 residues (start analysis with the sliding window length = 1 and increase). When treating alpha helix or beta structural periodicity, it is also advisable to have a sliding window of length from 3 to 20 residues. This is because alpha-helices and beta-strands are about this long. The user can also specify the type of analysis: SACC/SADC, multiple regression or neural networks analysis. The learning set includes all the proteins whose activities in the input data are represented as numbers. The proteins, whose activities in the

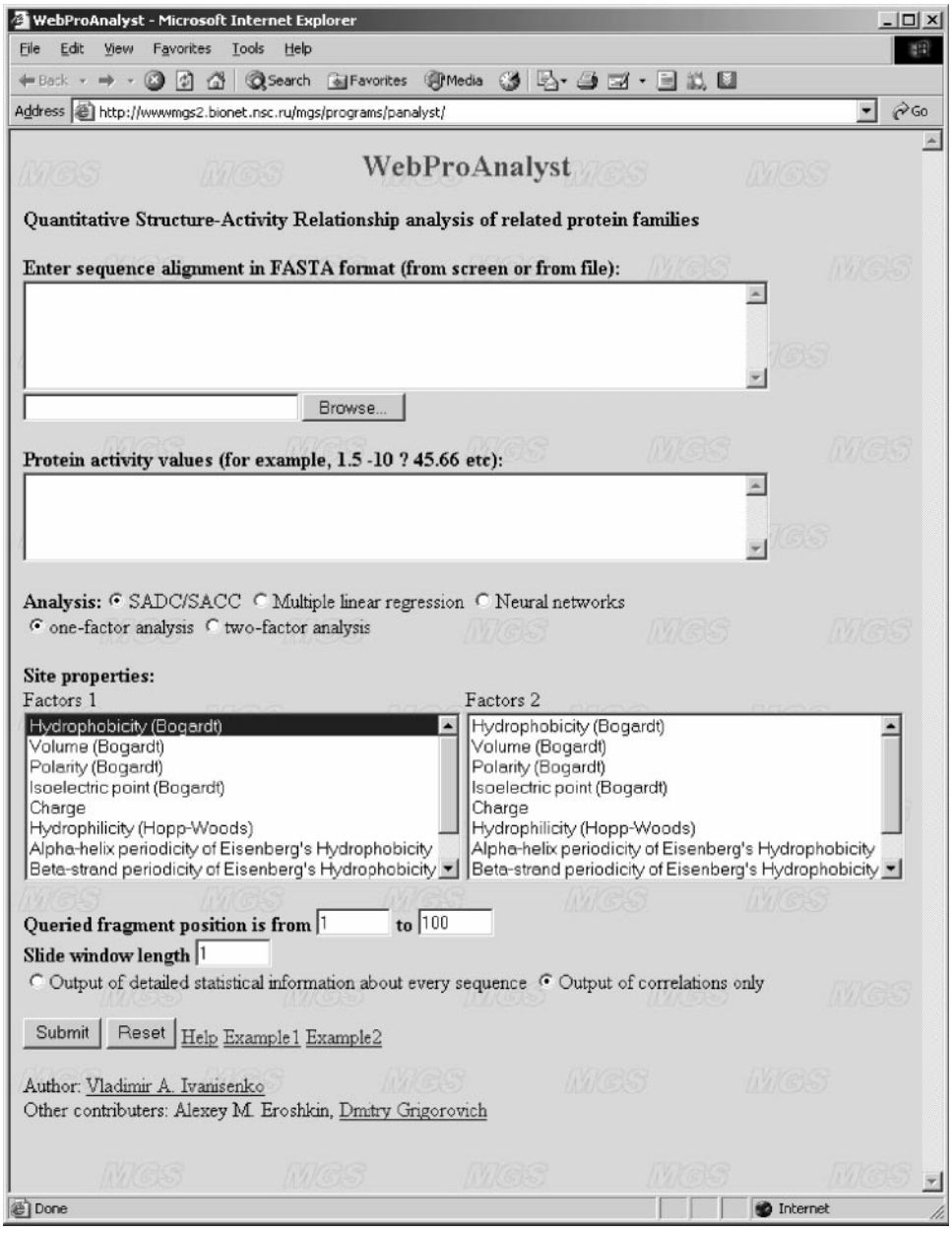

Figure 1. WebProAnalyst web page. The input data are the multiple alignment and protein activities. The output data are the models that relate the structure to protein activities. Depending on the parameters assigned by the user, building of the models is based either on multiple linear regression analysis, neural networks or SADC/SACC. The parameters include analysis type, site physicochemical properties, queried fragment boundaries and sliding window length.

input data have a question mark, are omitted from learning. The program provides automated calculation of the activities of the marked protein using the built models. Thus, a question mark is needed to warn that the activities of proteins of unknown function have to be predicted. The output contains the values for the predicted and measured activities, sequences and physicochemical characteristics of the analysed sites for every protein. For multiple linear regression and neural networks analysis, the values for the correlation between the predicted and the measured activities are given. Regression analysis also includes the regression equation and estimates of the significance of the parameters.

THE M2 PROTEIN OF INFLUENZA A VIRUSES

The M2 protein family of influenza A viruses was analysed to illustrate how the WebProAnalyst is applied in search of the physicochemical factors affecting the acquirement of viral resistance to amantadine. The M2 protein of influenza A viruses forms a proton channel involved in modifying the virion and the *trans*-Golgi pH during infection (13). The M2 protein ion channel activity is specifically blocked by the anti-influenza drug amantadine. The resistance of influenza A viruses to amantadine is caused by mutations in the transmembrane domain of the M2 protein (14). The M2 protein

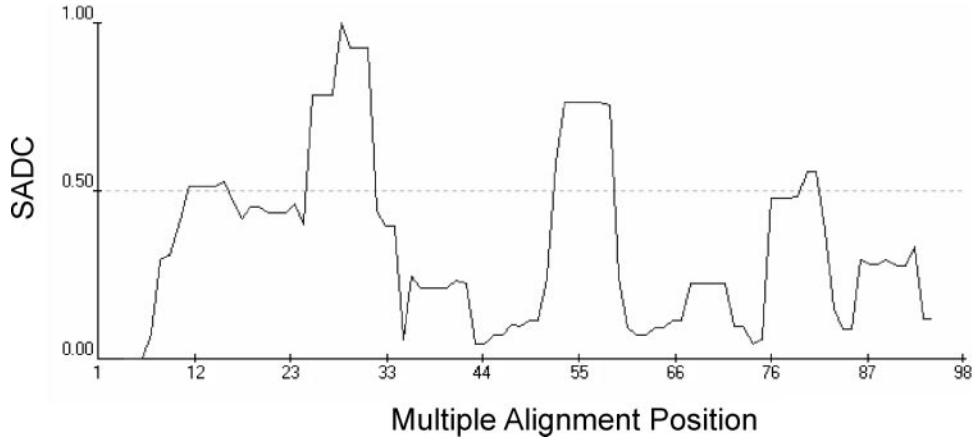

Figure 2. A SADC correlation profile calculated for the M2 protein of influenza A viruses using a seven residue sliding window. In the 25–31 region, which is involved in binding to amantadine, the correlation between amino acid substitutions and acquirement of virus resistance to amantadine is 100%.

sequences of influenza A virus derived from the amantadine resistant (21 proteins) and susceptible (30 proteins) strains were kindly provided by O. I. Kiselev (The Influenza Institute, St Petersburg, Russia). The M2 protein activity values for the amantadine resistant strain were assigned the value of 1 and the susceptible of 0.

Scanning with a seven residues long sliding window demonstrated a correlation between the influenza A viral drug resistance and the amino acid substitutions in the 25–31 region (the SADC being 100%, see Figure 2). Using neural networks and multiple linear regression analysis, we established that viral drug resistance correlates with the helix hydrophobic moment in this particular region ($R = 0.99$). Thus, substitutions that increase the values for the hydrophobic moment occur most frequently among the resistant strains of the virus. This means that the amphipathicity of the alpha-helix in the region is stronger in the drug resistant than the susceptible strains. There is abundant evidence indicating that the region we identified is important for the inhibition of viral infection with amantadine. Earlier, we have examined the role of the secondary structure of this region in the formation of viral resistance to rimantadine and deitiforin (15). In the presence of lipids, amantadine penetrates into the lipid membrane and interacts with M2 at specific sites, as neutron diffraction experiments have demonstrated previously (16). Amantadine is found near the centre of the bilayer, thereby suggesting the existence of an interaction site with Val-27 and Ser-31. Computationally, a minimum in the energy profile along the pore axis was revealed for amantadine in this region (17).

Another lower peak in the vicinity of residue 52 is distinguishable in the correlation profile (Figure 2). According to the secondary structure prediction, residues 46–62 form a strongly amphipathic helix that is possibly associated with the hydrophobic/hydrophilic interfacial region of the membrane (18). The correlations we established support the presumed important contribution of the region in the neighbourhood of residue 52 to the formation of the resistance of influenza A virus to amantadine. Here, we are not considering other peaks, e.g. around residues 80 and 15, that are significantly lower than the major peak. However, the appearance of these peaks in the SADC profile is suggestive, raising the question of whether the residues located in their regions may possibly be of importance in the formation of viral resistance; these particular residues may perhaps have an allosteric effect on the amantadine binding site. The residues in the region surrounding position 15 are along the border of the external and the transmembrane domains (13), and they may be involved in the early contacts between amantadine and M2 protein.

Table 1. Amino acid sequences and anti-*Candida* activity of histatin analogues (19)

Peptide	Activity IC_{50} (μM)	ln (IC_{50})	Sequence[a]
dh-5	4.1	1.41	KRKFHEKHHSHRGY
dh13L	5.2	1.64	K..L.....
dh15K	2.1	0.74	....K.....
dh17L	3.0	1.09	.....L.....
dh18L	3.0	1.09	.....L.....
dh18K	2.6	0.95	.....K.....
dh19K	2.5	0.91	.....K.....
dh21F	2.9	1.06	.....F...
dh23K	2.9	1.06	.....K.
dhvar1	0.6	−0.51	..L.K.LKF.L.K.
dhvar2	0.8	−0.22	..L.K.LLF.L.K.
dCysSN	114	4.73	SSPGKPPRLVG.P

[a]The matches between amino acids of peptides and peptide dh-5 are indicated by dots.

ANTIMICROBIAL PEPTIDES

Histatins have been described as histidine-rich cationic peptides, 7–38 amino acid residues in length with a strong killing effect *in vitro* on *Candida albicans*. Downstream of the C-terminal fungicidal domain of histatin 5 (residues 11–24) substitution analogues were synthesized and their antimicrobial activities were measured (19). We applied WebProAnalyst to study the effect of amphipathicity of the peptide on candidacidal activity. Table 1 gives a compilation of the amino acid sequences and their activities. The sequence scan using a sliding window of four residues allowed the detection of the most clear-cut correlations. The correlation between the helical hydrophobic moment in the region covering 3–6 residues and the antimicrobial activity was −0.97 (Figure 3). The stronger is the hydrophobic moment, the higher is the antimicrobial

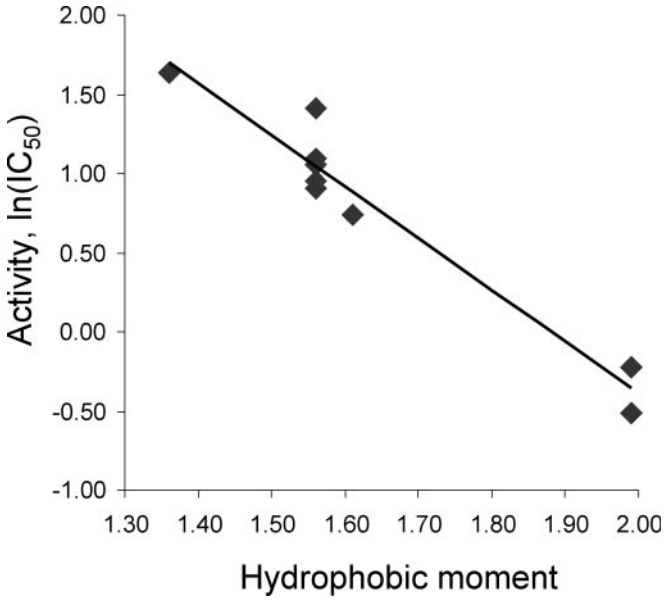

Figure 3. Linear relationship between the hydrophobic moment and the anti-*Candida* activity of histatin analogues. The correlation between the helical hydrophobic moment in the region covering the residues 3–6 and the antimicrobial activity is −0.97.

activity of the peptide. We did not use the dCysSN peptide for learning, consequently, its activity in the input data was marked with '?'. The neural networks predicted ln(IC_{50}) for this peptide to be 1.9 and the multiple linear regression analysis predicted a value of 5.8. These values are consistent with those for the lowest antimicrobial activity. The peptide is, indeed, devoid of activity and it has served as a negative control (19).

FUTURE DEVELOPMENTS

The idea behind all is to make WebProAnalyst workable for automated database search for sequences of such proteins whose predicted activities meet specific requirements. This would be reasonably achieved by integrating the WebProAnalyst and PSIBLAST programs. The other idea is to create a new WebProAnalyst module for the prediction of mutations that alter the protein activity in a targeted manner; the module would also estimate the effect of mutations on the structural integrity of a protein (whether they distort or break down its spatial structure). We further intend to integrate WebProAnalyst with PDBSiteScan, a program developed for the recognition of the functional protein sites (20,21). This would enable to make a more comprehensive annotation of protein structure–activity information. This might hopefully broaden the protein structure–activity realms of WebProAnalyst. The WebProAnalyst may be useful in a computational resolution of proteomics dilemmas, activity-based protein profiling is one (22,23).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Anna Fadeeva for translating the manuscript from Russian into English. The work was partly supported by the RFBR (grants 05-04-49283 and 03-04-48506-a), the SBRAS (Integration Project no. 119), the CRDF (Rup2-2629-NO-04, BRHE program NO-008-X1) and the Russian Ministry of Education and Science 'Development of the Scientific Potential Capacity of the Higher School' (Project no. 8274). Funding to pay the Open Access publication charges for this article was provided by the Institute of Cytology and Genetics SBRAS.

Conflict of interest statement. None declared.

REFERENCES

1. Hughes,T.R., Robinson,M.D., Mitsakakis,N. and Johnston,M. (2004) The promise of functional genomics: completing the encyclopedia of a cell. *Curr. Opin. Microbiol*., **7**, 546–554.
2. Mihalek,I., Res,I. and Lichtarge,O. (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol*., **336**, 1265–1282.
3. Livingstone,C.D. and Barton,G.J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci*., **9**, 745–756.
4. Kalinina,O.V., Novichkov,P.S., Mironov,A.A., Gelfand,M.S. and Rakhmaninova,A.B. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res*., **32**, W424–W428.
5. Lejon,T., Strom,M.B. and Svendsen,J.S. (2001) Antibiotic activity of pentadecapeptides modelled from amino acid descriptors. *J. Pept. Sci*., **7**, 74–81.
6. Sandberg,M., Eriksson,L., Jonsson,J., Sjostrom,M. and Wold,S. (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem*., **41**, 2481–2491.
7. Norinder,U., Rivera,C. and Unden,A. (1997) A quantitative structure–activity relationship study of some substance P-related peptides. A multivariate approach using PLS and variable selection. *J. Pept. Res*., **49**, 155–162.
8. Ivanisenko,V.A. and Eroshkin,A.M. (1997) Search for sites with functionally important substitutions in sets of related or mutant protein. *Mol. Biol. (Moskow)*, **31**, 749–755.
9. Eroshkin,A.M., Fomin,V.I., Zhilkin,P.A., Ivanisenko,V.A. and Kondrakhin,Y.V. (1995) PROANAL version 2: multifunctional program for analysis of multiple protein sequence alignments and studying structure-activity relationships in protein families. *Comput. Appl. Biosci*., **11**, 39–44.
10. Eroshkin,A.M., Zhilkin,P.A. and Fomin,V.I. (1993) Algorithm and computer program Pro__Anal for analysis of relationship between structure and activity in a family of proteins or peptides. *Comput. Appl. Biosci*., **9**, 491–497.
11. Draper,N.R. and Smith,H. (1998) *Applied Regression Analysis, 3rd edn*. John Wiley and Sons, NY.
12. Rumelhart,D.E., Hinton,G.E. and Williams,R.J. (1986) Learning internal representations by error propagation. In Rumelhart,D.E. and McClelland,J.L. (eds), *Parallel Distributed Processing*. MIT Press, CambridgeVol. 1, pp. 318–362.
13. Hay,A.J. (1992) The action of adamantanamines against influenza A viruses: inhibition of the M2 ion channel protein. *Semin. Virol*., **3**, 21–30.
14. Chizhmakov,I.V., Geraghty,F.M., Ogden,D.C., Hayhurst,A., Antoniou,M. and Hay,A.J. (1996) Selective proton permeability and pH regulation of the influenza virus M2 channel expressed in mouse erythroleukaemia cells. *J. Physiol*., **494**, 329–336.
15. Kiselev,O.I., Mishin,V.P., Eroshkin,V.I., Kozeletskaia,K.N., Usova,E.V., Rudenko,V.I. and Chupakhin,O.N. (1994) Secondary structure of the M2 protein of influenza type A virus and its role in forming resistance to rimantadine and deitiforin. *Mol. Biol. (Moskow)*, **28**, 1009–1013.
16. Duff,K.C., Gilchrist,P.J., Saxena,A.M. and Bradshaw,J.P. (1994) Neutron-diffraction reveals the site of amantadine blockade in the influenza-A M2 ion-channel. *Virology*, **202**, 287–293.

17. Sansom,M.S.P. and Kerr,I.D. (1993) Influenza virus M2 protein: a molecular modelling study of the ion channel. *Protein Eng*., **6**, 65–74.
18. Tian,C., Gao,P.F., Pinto,L.H., Lamb,R.A. and Cross,T.A. (2003) Initial structural and dynamic characterization of the M2 protein transmembrane and amphipathic helices in lipid bilayers. *Protein Sci*., **12**, 2597–2605.
19. Helmerhorst,E.J., Van't Hof,W., Veerman,E.C., Simoons-Smit,I. and Nieuw Amerongen,A.V. (1997) Synthetic histatin analogues with broad-spectrum antimicrobial activity. *Biochem. J*., **326**, 39–45.
20. Ivanisenko,V.A., Pintus,S.S., Grigorovich,D.A. and Kolchanov,NA. (2005) PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res*., **33**, D183–D187.
21. Ivanisenko,V.A., Pintus,S.S., Grigorovich,D.A. and Kolchanov,N.A. (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res*., **32**, W549–W554.
22. Fliri,A.F., Loging,W.T., Thadeio,P.F. and Volkmann,R.A. (2005) Biological spectra analysis: linking biological activity profiles to molecular structure. *Proc. Natl Acad. Sci. USA*, **102**, 261–266.
23. Liu,Y., Patricelli,M.P. and Cravatt,B.F. (1999) Activity-based protein profiling: the serine hydrolases. *Proc. Natl Acad. Sci. USA*, **96**, 14694–14699.