Mol. Cells *28*, 195-200, September 30, 2009
DOI/10.1007/s10059-009-0122-z

Molecules and Cells

A Computational Approach for the Classification of Protein Tyrosine Kinases

Hyun-Chul Park[1], Hae-Seok Eo[2], and Won Kim[1,3,*]

Protein tyrosine kinases (PTKs) play a central role in the modulation of a wide variety of cellular events such as differentiation, proliferation and metabolism, and their unregulated activation can lead to various diseases including cancer and diabetes. PTKs represent a diverse family of proteins including both receptor tyrosine kinases (RTKs) and non-receptor tyrosine kinases (NRTKs). Due to the diversity and important cellular roles of PTKs, accurate classification methods are required to better understand and differentiate different PTKs. In addition, PTKs have become important targets for drugs, providing a further need to develop novel methods to accurately classify this set of important biological molecules. Here, we introduce a novel statistical model for the classification of PTKs that is based on their structural features. The approach allows for both the recognition of PTKs and the classification of RTKs into their subfamilies. This novel approach had an overall accuracy of 98.5% for the identification of PTKs, and 99.3% for the classification of RTKs.

INTRODUCTION

Protein tyrosine kinases (PTKs) are a large multigene family found in most of eukaryotic organisms, and they play an important role in cellular processes such as metabolism, migration, survival, proliferation, and differentiation (Hanks and Hunter, 1995; Hunter, 1991). Unregulated activation of PTKs can cause various diseases, including cancer, diabetes and atherosclerosis (Dean et al., 1985; Robertson et al., 2000; Ullrich et al., 1985). On account of the large diversity within the family and their important roles in cellular processes, the field of drug development has focused on utilizing PTKs as target molecules (Mendelsohn and Baselga, 2000; Sridhar et al., 2000; Tidow et al., 2004).

PTKs contain an evolutionarily conserved catalytic kinase domain capable of phosphorylating substrate proteins on tyrosine residues and can be divided into two classes according to the presence of a transmembrane (TM) domain: receptor tyrosine kinases (RTKs) and non-receptor tyrosine kinases (NRTKs). RTKs are composed of a single transmembrane domain, an extracellular N-terminal region, which binds to ligands, and an intracellular C-terminal region responsible for the kinase activity. NRTKs consist of a single tyrosine kinase domain without a transmembrane domain, and several domains capable of protein-protein interactions (Fig. 1). Most characterized RTKs and NRTKs, which are encoded in the human genome, can be subdivided into 20 families (including 1 pseudogene family) and 10 subfamilies according to their specific domain structures (Hubbard and Till, 2000; Neet and Hunter, 1996; Robinson et al., 2000). Many of the PTKs in human and other model organisms have been identified by experimental methods, and several research groups have characterized the function of PTKs (Chiarugi, 2008; Chu et al., 1996; Kong et al., 2008; Partanen et al., 1996). However, it is costly and time-consuming to assign a function using these experimental methods because of their structural complexity. Here, we describe a novel approach for identifying and classifying PTKs in order to overcome the limitations described above. To enhance the diagnostic power of the approach, the fingerprint method, which uses specific subdomains in catalytic kinase domains and transmembrane domains, and profile hidden Markov models (HMMs) were combined (Eo et al., 2007). The performance of this novel classification method was evaluated by cross-validation and showed a high discriminative potency in the identification and classification of PTKs.

MATERIALS AND METHODS

The method of this study was composed of 4 steps; data collection, domain extraction, construction of profile HMM library and validation (Fig. 2). The dataset was acquired from databases, and divided into training set and test set for building and evaluating the model, respectively. The motif and TM domain which will be used to build the HMMs library were extracted with the published paper and ontology information. After a process of alignment of each domain sequence was performed, profile HMM library based on motif and TM was constructed using HMMER software. We tried to identify PTKs and to classify RTKs using test set, and evaluated the result using two cross validation method, holdout and *n*-fold cross validation.

Datasets

The PTKs dataset was obtained for building and evaluating the model from SwissProt (release 56.2, http://br.expasy.org/sprot/;

[1]Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea, [2]School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Korea, [3]School of Biological Sciences, College of National Sciences, Seoul National University, Seoul 151-742, Korea
*Correspondence: wonkim@plaza.snu.ac.kr

Received July 8, 2009; accepted July 20, 2009; published online September 4, 2009

Keywords: classification, motifs and transmembrane domain, profile hidden Markov model, protein tyrosine kinase, receptor tyrosine kinase

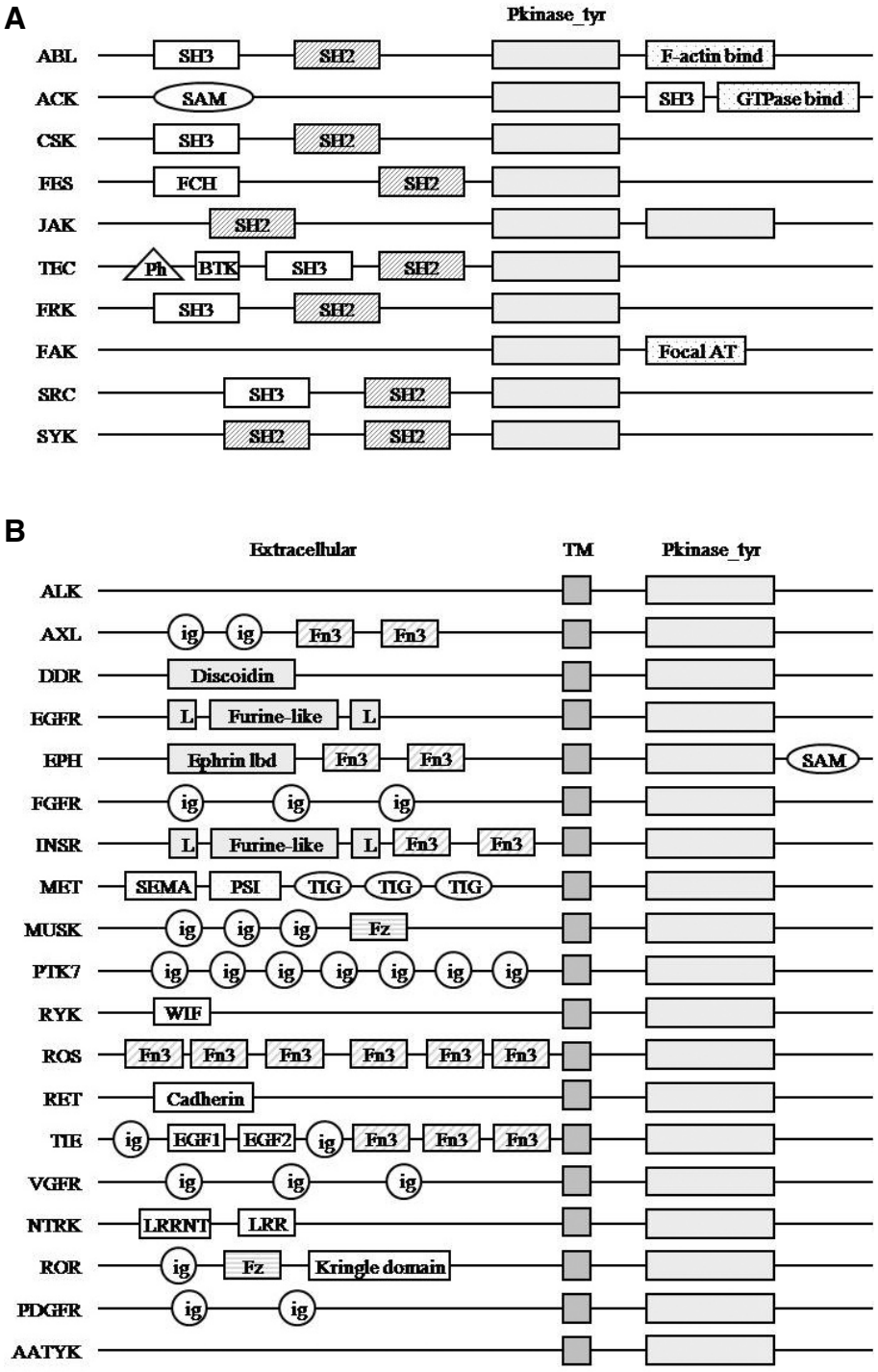

Fig. 1. A schematic structure of the human protein tyrosine kinase families redrawn from Robinson et al. (2000). The structure was obtained from the Pfam database (version 21.0).

Boeckmann et al., 2003) using the following queries: description [Tyrosine] and organism [Mammalia] in the field name. The 89 human PTKs sequences were obtained (Robinson et al., 2000) and used to construct the motif-based model (Supplementary Table 1). Of the PTKs sequences, 186 RTKs sequences were used in order to train a TM-based model for classifying RTKs into their subfamilies. The test set for evaluating the method consisted of 624 protein kinase group sequences and 764 RTK sequences obtained through a BLAST search. The protein kinase groups were obtained from the 'Human Kinome' database (http://kinase.com/human/kinome/; Manning et al., 2002). This dataset is accessible from : http://147.47.216.128.

Motif and TM domain identification

The catalytic domain of PTKs is an evolutionary conserved region that has particular motifs which can discriminate PTKs. We identified 5 specific motifs that only the tyrosine kinase group had from published papers and confirmed the presence of the motifs within the catalytic domain. (Hanks and Quinn, 1991; Hanks et al., 1988; Pearson and Kemp, 1991; Table 1).

Table 1. The 5 specific motifs of protein tyrosine kinase and log-odd scores for the consensus and exceptional sequence

No.	Motif	Consensus / Exceptional	Probability	Log-odd score
1	HRDLX[1]ARN	HRDLAARN	0.481	23.23
		HSS(G)I(V)SS(G,V)HS(H,K)	8×10^{-14}	-6.2
2	PXXWXAPE	PIKWMAPE	0.062	21.18
		A(I)A(F)LWL(V)SWE	2.4×10^{-11}	-0.49
3	SDVWSXG	SDVWSFS	0.417	20.096
		G(M)NS(M)WGSA	4×10^{-12}	-5.274
4	KXXDFG	KISDFG	0.149	16.071
		QLTA(R,G)NA	2.4×10^{-11}	-6.471
5	CW	CW	0.921	5.9
		TL	2.1×10^{-4}	-2.478

[1]any amino acid

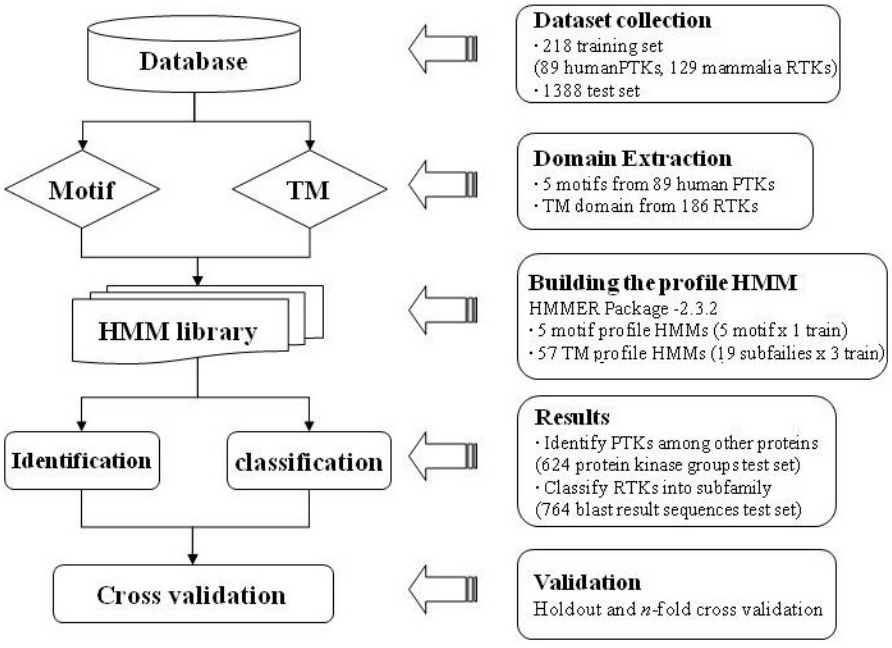

Fig. 2. Flowchart of PTK classification and identification

We also extracted the transmembrane (TM) region of 186 RTKs sequences (57 human and 129 mammalia) based on the description of 'TRANSMEM' in the FT line (Supplementary Table 2).

Construction of the profile HMM

To make the motif-based profile HMM library, a multiple sequence alignment was performed for the catalytic domains of PTKs using ClustalW1.83 (Supplementary Fig. 1.). The 5 motifs region was parsed from the aligned catalytic domains, and each aligned motif sequence was used as an insert data to program. Moreover, the TM sequence data was randomly divided into 3 training sets to perform *n*-fold cross validation (Supplementary Table 3). The separate sequence was aligned, and a TM-based profile HMM library per subfamily was constructed. When there is only one sequence in subfamily such as MUSK, RET and ROR, we ran the program with just the sequence.

We used the HMMER software ver. 2.3.2 (Eddy, 1998) for construction of profile HMMs. The profile HMMs were built using the *hmmbuild* program of the HMMER package with option *-A,* and calibrated using the *hmmcalibrate* program in order to improve the sensitivity of the database search. The 5 motif-based and 57 (19 subfamilies x 3 training set) TM-based profile HMM libraries were constructed in this study. The *hmmpfam* program reads the query sequences and searches the matched similar sequences against a profile HMM database. We examined test set sequences as a query sequence against both motif and TM-based profile HMM libraries.

Performance estimation

In order to estimate the performance and accuracy of our profile HMM libraries, two validation methods were adopted in this study; a holdout and an *n*-fold cross validation. Since a holdout validation is the simplest validation method, it was used to identify PTK groups from a motif-based profile HMMs. An *n*-fold cross validation was used in order to avoid an overestimation of

Table 2. Prediction of the entire protein kinase family

Group	No. of predictions	Error rate	Accuracy
AGC	4 / 69	0.057 (FP)	0.943
CaMK	2 / 113	0.017 (FP)	0.983
CK1	0 / 17	0 (FN)	1.00
CMGC	0 / 73	0 (FN)	1.00
RGC	0 / 8	0 (FN)	1.00
STE	0 / 53	0 (FN)	1.00
TK	**95 / 95**	**0 (TP)**	**1.00**
TKL	3 / 49	0.061 (FP)	0.939
Atypical	0 / 43	0 (FN)	1.00
Others	0 / 104	0 (FN)	1.00
Total	9 / 624	0.014	0.986

the correct classification rate. We adopted *n*-fold cross validation to evaluate classification of RTKs using a TM-based profile HMMs. Initially, the sequences of all subfamilies of RTKs were randomly divided into *n* subsets. Subsequently, one of the *n* subsets was used as a test set and the remaining *n*-1 subsets were used as the training set in the model. This procedure was repeated *n* times, and the error rate was calculated in each step. In the final step, the true error rate was calculated as the average of the individual error rates, E_i

$$E = \frac{1}{n}\sum_{i=1}^{n} E_i$$

The test set was divided into true positive (TP), true negative (TN), false positive (FP), and false negative (FN) according to whether the PTKs belong to their own family. The performance of this approach was then estimated by the accuracy of classification.

RESULTS AND DISCUSSION

Evaluation of motif-based profile HMM

The probabilities of each position of 5 specific motifs were calculated in order to build the profile HMMs using consensus sequences. The probability and log-odd score of each motif were firstly calculated using consensus and exceptional sequences to examine the accuracy of the model (Table 1). The probabilities of 5 motifs were calculated by multiplying the probability of the transition from the previous state to the next state and the probability of occurrence at each position (here, transition probability is 1.0). However, since probability depends on the length of the sequence, it was not suitable to use probability as a score. A more appropriate score was the log-odd score, which is the logarithm of the probability of the sequence divided by the probability according to a null model. The log-odd score was calculated using the following equation:

$$\text{Log-odd score for sequence } S = \log_2 \frac{p(s)}{p(null)^L} = \log_2 P(S) - L \log_2 P(null)$$

If the log-odd score was high, the sequence was suitable to the model. However, if it was negative, the sequence was close to the null model (Krogh, 1998).

Identification of PTK

To estimate the accuracy, we subjected the model to 624 test set sequences. Since the motif-based model was constructed using 89 human tyrosine kinase sequences, 180 non-human PTK sequences were first applied in order to evaluate whether the model was appropriate for classification of PTKs in species other than human (data are not provided). From this analysis, we found that all non-human sequences were correctly classified as belonging to the approriate tyrosine kinase family. Moreover, 95 tyrosine kinases in 624 sequences were identified with 100% accuracy (Bold in Table 2). However, some sequences that did not belong to the tyrosine kinase family were falsely predicted under the e-value threshold. In these cases, they belonged to other kinase families: cAMP-depen-dent protein kinase/protein kinase G/protein kinase C (AGC), Ca^{2+}/calmodulin-dependent kinase (CaMK), and tyrosine kinase- like (TKL) group. The average accuracy in identifying PTKs among the entire protein kinase groups was more than 98%. This measurement was adjusted using an e-value threshold of 0.026. These combined results imply that the specific motifs within the catalytic domain have sufficient information to discriminate between other protein kinases with high accuracy.

Table 3. True error rate of *n*-fold cross-validation according to the e-value ($n = 3$)

E-value	True error rate	Correction rate
0.1	0.251	0.749
0.01	0.086	0.914
0.026	0.048	0.952

Classification of the receptor tyrosine kinase subfamilies

The transmembrane region is the most essential component of the receptor-type membrane protein. The transmembrane domain has a high sequence similarity between subtypes within the same subfamily; however, it has considerably low similarity between subfamilies. We classified RTKs into their own family using the differences between sequence similarity. In this step, the estimation of the error rate and correction rate was performed with 3-fold cross validation according to 3 e-value scores, and the results are shown in Table 3.

The process of calculating the error rate is described in "Materials and Methods." The model, which was based on the TM domains, showed the highest accuracy (95%) when an e-value of 0.026 was used. Model estimation was performed using a test set of 624 sequences and 764 RTK sequences (Dataset section in "Materials and Methods"). Of the 95 PTK sequences shown in Table 2, 59 sequences were RTKs, among which 57 were correctly classified into their subfamily. Two sequences that were regarded as true negatives were the pseudogenes of RTKs, namely, *FLT1ps* and *Tyro3ps*. Moreover, for the 764 RTK sequences, most of the subfamilies were correctly classified with an accuracy of 100%, although the model did not accurately predict 6 sequences of 3 subfamilies (EPH, NTRK and PDGFR) (bold in Table 4).

When 624 sequences of human protein kinases were tested using both the motif- and TM-based profile HMM libraries, the entire performance of this method, i.e., the ability to classify PTKs, yielded an accuracy of 98.5% and 99.3%, respectively (Table 5).

In this study, all protein tyrosine kinase groups were identified by three motifs; HRDLXARN (subdomain VI), KXXDFG (subdomain VII), SDVWSXG (subdomain IX). Although the motif, PXXWXAPE (subdomain VIII), is a conserved motif in tyrosine kinases, it does not contribute to the specificity of tyrosine (Carrera et al., 1994).

Table 4. Error rate and accuracy of each subfamily of RTKs

Subfamily	No. of detection	Error rate	Accuracy
ALK	33 / 33	0 (TP)	1.00
AXL	38 / 38	0 (TP)	1.00
DDR	14 / 14	0 (TP)	1.00
EGFR	56 / 56	0 (TP)	1.00
EPH	**76 / 80 (-4)**	**0.05 (TN)**	**0.95**
FGFR	117 / 117	0 (TP)	1.00
INSR	49 / 49	0 (TP)	1.00
MET	5 / 5	0 (TP)	1.00
MUSK	20 / 20	0 (TP)	1.00
PTK7	20 / 20	0 (TP)	1.00
RYK	29 / 29	0 (TP)	1.00
ROS	16 / 16	0 (TP)	1.00
RET	17 / 17	0 (TP)	1.00
TIE	35 / 35	0 (TP)	1.00
VGFR	79 / 79	0 (TP)	1.00
NTRK	**17 / 18 (-1)**	**0.055 (TN)**	**0.945**
ROR	82 / 82	0 (TP)	1.00
PDGFR	**18 / 19 (-1)**	**0.052 (TN)**	**0.948**
AATYK	37 / 37	0 (TP)	1.00
Total	758 / 764	0.008	0.992

Table 5. The overall prediction according to each library

HMM library	Se[1]	Sp[2]	Acc[3]	Mcc[4]
Motif-based profile HMM library	1	0.982	0.985	0.947
TM-based profile HMM library	0.966	0.996	0.993	0.962

[1]sensitivity
[2]specificity
[3]accuracy
[4]Matthews correlation coefficient

We designed an approach algorithm that recognizes and classifies the tyrosine kinase group using their specific motifs and the TM domains. This approach was successful because the HMM, which was based on motifs, had a high accuracy in the database searches, and the TM domains contained enough information to discriminate among functional families (Grundy et al., 1997; Sadka and Linial., 2005). As mentioned in the previous section, '*Identification of PTK*', this method yielded a considerably high accuracy of more than 98% in the classification of RTK subfamilies as well as the entire protein tyrosine kinase family. In a report that used bioinformatics approaches to classify GPCRs, diagnostic performance was found to be enhanced when characteristic fingerprints based on motifs were used, and the use of profile HMM for protein domains performed well in the detection of remote family members among large superfamilies (Gaulton and Attwood, 2003). Furthermore, the profile HMMs that utilize the fingerprint approach showed higher accuracy than methods using other algorithms (Eo et al., 2007).

Some pairwise alignment tools, BLAST and FASTA, are simple and frequently used methods that identify and classify the novel proteins and genes using a similarity of their sequence. However, these programs have a few limitations such as confidence of results, return of different results and determination of level of sequence identity. To overcome the limitations of pairwise tools, many protein databases and methods that utilize computational algorithms have been developed (Gaulton and Attwood, 2003).

In several recent biological studies, computational methods have been used to predict experimental results (Karchin et al., 2002; Melodelima et al., 2006; Sgourakis et al., 2005a; 2005b; Wang et al., 2005; Weinert and Lopes, 2004). A statistical estimation can prove to be quite useful since it can identify and reduce the number candidate genes in an exhaustive biological analysis (Mitrophanov and Borodovsky, 2006). The present method will be useful to identify unknown genes and hypothetical proteins in ongoing sequencing projects of other species within the metazoan, and it can be used for the classification of other protein kinase groups. Additionally, the accurate classification of PTKs can provide a better understanding of their structure and functions and play an important role in the field of drug development.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

This work was in part supported by the second stage of the Brain Korea 21 Project in 2008, a grant from the Eco-Technopia 21 Project funded by the Ministry of Environment, and a grant from the Marine Biotechnology Programme funded by the Ministry of Land, Transport and Maritime Affairs of Korean government.

REFERENCES

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. (2003). The SWISS-PROT protein knowledge-base and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365-370.

Carrera, A.C., Borlado, L.R., Roberts, T.M., and Martinez-A.,C. (1994). Tyrosine kinase specific motif at subdomain VIII does not confer specificity for tyrosine. Biochem. Biophys. Res. Commun. *205*, 1114-1120.

Chiarugi, P. (2008). Src redox regulation: There is more than meets the eye. Mol. Cells *26*, 329-337.

Chu, D.H., Spits, H., Peyron, J.-F., Rowley, R.B., Bolen, J.B., and Weiss, A. (1996). The Syk protein tyrosine kinase can function independently of CD45 or Lck in T cell antigen receptor signaling. EMBO J. *15*, 6251-6261

Dean, M., Park, M., Le Beau, M.M., Robins, T.S., Diaz, M.O., Rowley, J.D., Blair, D.G., and Vande Woude, G.F. (1985). The human *met* oncogene is related to the tyrosine kinase oncogenes. Nature *318*, 385-388.

Eddy, S.R. (1998). Profile hidden markov model. Bioinformatics *14*, 755-763.

Eo, H.S., Choi, J.P., Noh, S.J., Hur, C.G., and Kim, W. (2007). A combined approach for the classification of G protein-coupled receptors and its application to detect GPCR splice variants. Comput. Biol. Chem. *31*, 246-256.

Gaulton, A., and Attwood, T.K. (2003). Bioinformatics approaches for the classification of G-protein-coupled receptors. Curr. Opin. Pharmacol. *3*, 114-120.

Grundy, W.N., Bailey, T.L., Elkan, C.P., and Baker, M.E. (1997). Meta-MEME: Motif-based hidden Markov models of protein families. Comput. Appl. Biosci. *13*, 397-406.

Hanks, S.K., and Hunter, T. (1995). The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB *9*, 576-579.

Hanks, S.K., and Quinn, A.M. (1991). Protein kinase catalytic domain sequence database : Identification of conserved features of primary structure and classification of family members. Methods Enzymol. *200*, 38-62.

Hanks, S.K., Quinn, A.M., and Hunter, T. (1988). The protein kinase family: conserved features and deduced phylogeny of the catalytic

domains. Science *241*, 42-52.
Hubbard, S.R., and Till, J.H. (2000). Protein tyrosine kinase structure and function. Ann. Rev. Biochem. *69*, 373-398.
Hunter, T. (1991). Protein kinase classification. Methods Enzymol. *200*, 3-37
Karchin, R., Karplus, K., and Haussler, D. (2002). Classifying G-protein coupled receptors with support vector machines. Bioinformatics *18*, 147-159.
Kong, A., Calleja, V., Leboucher, P., Harris, A., Parker, P.J., and Larijani, B. (2008). HER2 oncogenic function escapes EGFR tyrosine kinase inhibitors via activation of alternative HER receptors in breast cancer cells. PLoS One *3*, e2881.
Krogh, A. (1998). Computational methods in molecular biology. In S.L., Salzberg, D.B., Searls, and S., Kasif, eds. (Amsterdam: Elsevier Science B.V.), pp. 45-64.
Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912-1934.
Melodelima, C., Gueguen, L., Piau, D., and Gautier, C. (2006). A computational prediction of isochores based on hidden Markov models. Gene *385*, 41-49.
Mendelsohn, J., and Baselga, J. (2000). The EGF receptor family as targets for cancer therapy. Oncogene *19*, 6550-6565.
Mitrophanov, A.Y., and Borodovsky, M. (2006). Statistical significance in biological sequence analysis. Brief Bioinform. *7*, 2-24.
Neet, K., and Hunter, T. (1996). Vertebrate non-receptor protein-tyrosine kinase families. Genes Cells *1*, 147-169.
Partanen, J., Puri, M.C., Schwartz, L., Fischer, K.D., Bernstein, A., and Rossant, J. (1996). Cell autonomous functions of the receptor tyrosine kinase TIE in a late phase of angiogenic capillary growth and endothelial cell survival during murine development. Development *122*, 3013-3021.
Pearson, R.B., and Kemp, B.E. (1991). Protein kinase phosphorylation site sequences and consensus specificity motifs: Tabulations Methods Enzymol. *200*, 62-81.
Robertson, S.C., Tynan, J.A., and Donoghue, D.J. (2000). RTK mutations and human syndromes. Trends Genet. *16*, 265-271.
Robinson, D.R., Wu, Y.-M., and Lin, S.-F. (2000). The protein tyrosine kinase family of the human genome. Oncogene *19*, 5548-5557.
Sadka, T., and Linial, M. (2005). Families of membranous proteins can be characterized by the amino acid composition of their transmembrane domains. Bioinformatics *21*, i378-i386.
Sgourakis, N.G., Bagos, P.G., and Hamodrakas, S.J. (2005a). Prediction of the coupling specificity of GPCRs to four families of G-proteins using hidden Markov models and artificial neural networks. Bioinformatics *21*, 4101-4106.
Sgourakis, N.G., Bagos, P.G., Papasaikas, P.K., and Hamodrakas, S.J. (2005b). A method for the prediction of GPCRs coupling specificity to G-proteins using refined profile hidden Markov models. BMC Bioinformatics *6*, 104.
Sridhar, R., Hanson-Painton, O., and Cooper, D.R. (2000). Protein kinases as therapeutic targets. Pharm. Res. *17*, 1345-1353.
Tidow, C.M., Schwable, J., Steffen, B., Tidow, N., Brandt, B., Becker, K., Bahr, E.S., Halfter, H., Vogt, U., Metzger, R., et al. (2004). High-throughput analysis of genome-wide receptor tyrosine kinase expression in human cancers identifies potential novel drug targets. Clin. Cancer Res. *10*, 1241-1249.
Ullrich, A., Bell, J.R., Chen, E.Y., Herrera, R., Petruzzelli, L.M., Dull, T.J., Gray, A., Coussens, L., Liao, T.-C., Tsubokawa, M., et al. (1985). Human insulin receptor and its relationship to the tyrosine kinase family of oncogenes. Nature *313*, 756-761.
Wang, Y-F., Chen, H., and Zhou, Y.-H. (2005). Prediction and classification of human G-protein coupled receptors based on support vector machines. Genomics Proteomics Bioinformatics *3*, 242-246.
Weinert, W.R., and Lopes, H.S. (2004). Neural networks for protein classification. Appl. Bioinformatics *3*, 41-48.