Comprehensive Analysis of Protein Kinase Domains

Myriam Passamano[(1)] , Nunzio D'Agostino[(1)], Andrea Caprera[(1)], Luciano Milanesi[(1)].

[(1)]Istituto Tecnologie Biomediche CNR, Via Fratelli Cervi 93, 20090 Segrate (MI)- Italy
luciano.milanesi@itb.cnr.it

Keywords. Kinase, Domains, MyKinDB Database .

Introduction

Eukaryotic protein kinase (ePKs) constitute one of the largest recognized protein families represented in the human genome and are important players in virtually every signaling pathway involved in normal development and disease. The key feature that distinguishes ePKs superfamily members from other proteins is the sequence of contiguous stretch of approximately 250 aminoacids that constitutes the catalytic domain [1-2] . Around half the human kinases contain other domains in addition to the catalytic domain, which often are involved in kinases regulation, interactions with other partners or subcellular localization [3]. Domains present one of the most useful levels at which to understand protein function and domain family-based analysis, so we developed an automated analysis system for studying domain statistic distribution of kinase superfamily.

Materials and Methods

Human kinase dataset: The complete set of human kinase was obtained from *KinBase*, the kinase database at SUGEN (http://198.202.68.14/kinbase). A specific program module has been developed to gain all 624 human protein kinase records into the mySQL relational database 'MyKinDB'. Throught a SQL query, then we identified and recorded into the *Kinase_pseudogenes* table all 106 pseudogenes.

Domains analysis: We used InterProScan package, version 3.1 from EBI on a Red Hat 8.0 Linux platform in order to use the following softwares: ScanRegExp, ProfileScan, FprintScan, HMMPfam, HMMPIR PIR, Superfamily. By using a program module, we extracted each individual kinase protein sequence in FASTA format from the *MyKinDB* database and inputed the result into InterProScan. The output, in XML format, has been parsed and the extracted information has been recorded into '*Domains*' table. On the other hand, information about the start and end domain position, have been recorded in seven different tables, each of these refer to each individual member database (i.e.: 'int_smart'). This automated analysis has demanded 56 hours to be complete on an Intel Pentium 4 CPU 1.80GHz, RAM 522Mb. In figure 1 the main steps of our analysis system are shown.

Results and Discussion

The high degree of functional diversification among the protein kinases is made possible by their ability to interact with large numbers of cellular proteins. These interactions are mediated through additional subunits or domains of the kinase that are regulatory or act as protein-interaction modules. The presence of these non-catalytic domains explains the high functional diversification among kinases and suggests alternative strategies of cellular processes regulation. Our goal has been the development of an InterProScan automated analysis to study domain statistic distribution of kinase superfamily. The analysis of the 518 protein kinase sequences recorded in *MyKinDB* database have lead us to the identification of 91 domains, 12 Repeats, 1 Binding site and 20 families according to the following Interpro cascade classification schema: Family, Domain, Repeat and Site. A majority (about 80%) of the human protein kinases contain at least one domain other than the catalytic kinase domain. *TEC*, *SRC*, *CSK* and *ABL*, for example, are TK families that contain *Src Homology 2 (SH2) and Src Homology 3 (SH3)* domains before the catalytic domain. Both are adaptor domains and are involved in the recruitment of proteins to their specific target. SH3 domain precedes and it is contiguous to the SH2 domain. We have also studied the peptide that binds together SH2 and catalytic domain. The length of this linker peptide is in average 20 amino-acids long. *Gonfloni et al.* revealed that *SH2-CatalyticDomain linker* is a critical player in the regulation activity of these tyrosine kinases families. They found that mutation of Tryptophan to Alanine clearly impairs kinases regulation and suggest that this Tryptophan residue may also have a structural role in agreement with its high degree of conservation among protein tyrosine kinases (fig 2) [4]. *TIE*, *AXL*, *ROS* and *InsR* families, instead, are TK families that contain *Fibronectin type III domain.* From 2 to 8 modules precede catalytic domain. Each module is approximately 100 amino acid long and contain different tandem repeats which are binding sites for DNA, heparin and the cell surface. Nearly all kinases belonging to AGC group contain protein kinase C-terminal domain. It is an accessory domain often found associated with catalytic domain. It is a calcium-activated and phospholipid-dependent domain. The remaining domains are grouped according to their substrate specificity: 53 kinases present domains linked to lipid signalling, 37 kinases with domains linked to GTPase signalling, 12 kinases present domains involved in calcium signalling, 8 kinases present domains able to target the protein to the cytoskeleton and 60 kinases present domain that interact with nucleic acids. In general, we have observed that members of the same kinase

family have the same domain composition, even if some domain shuffling is seen. This imply a coevolution of catalytic domain and associated domains and suggest that the cross-talk amog different signaling pathways, although complicated, is achieved using a limited number of modules. The occurrence of combinations of these modules would narrow the range of functional diversity.

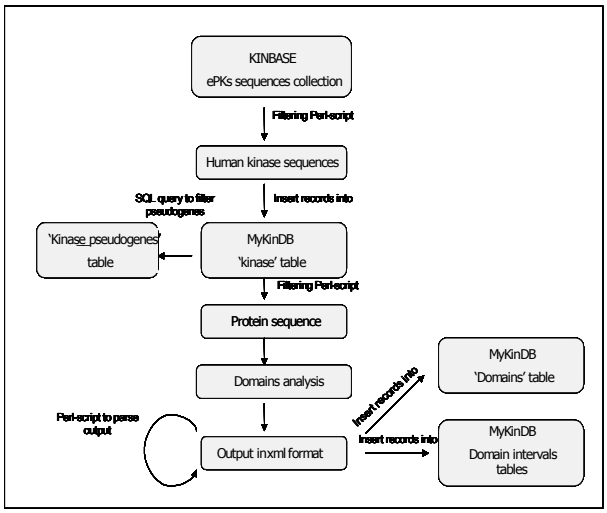

Fig 1. Step by step schema of our analysis system .

```
ABL.AA_220_241  ----RNKPTVYGVSP-NYDKWEM-ERTD
ARG.AA_231_251  -----NKPTVYGVSP-IHDKWEM-ERTD
FGR.AA_242_262  ---TIMKPQTLGLA---KDAWEI-SRSS
BTK.AA_380_401  --QNKNAPSTAGLG---YGSWEI-DPKD
ITK.AA_340_362  -FGRQKAPVTAGLR---YGKWVI-DPSE
FYN.AA_249_270  -----GMPRLTDLSVKTKDVWEI-PRES
LYN.AA_227_246  -----ISPK--PQKPWDKDAWEI-PRES
CSK.AA_172_194  -VMEGTVAAQDEFY---RSGWAL-NMKE
CTK.AA_214_234  ---HGTKSAEEELA---RAGWLL-NLQH
HCK.AA_243_261  ----SSKPQ----KPWEKDAWEI-PRES
BLK.AA_224_240  --------APQNP--WAQDEWEI-PRQS
FRK.AA_209_233  --LKIQVPAPFDLSYKTVDQWEI-DRNS
LCK.AA_227_244  -----QKPQ----KPWWEDEWEV-PRET
BRK.AA_173_190  -----HEPEPL---P-HWDDWER-PREE
BMX.AA_398_421  STKANKVPDSVSLG---NGIWEL-KREE
SRC.AA_251_269  -----SKPQTQGLA---KDAWEI-PRES
TEC.AA_346_369  SVKGKNAPTTAGFS---YEKWEINPSE-
TXK.AA_247_270  GLMGSCLPATAGFS---YEKWEIDPSE-
YES.AA_258_276  -----VKPQTQGLA---KDAWEI-PRES
SRM.AA_213_229  ------MPQKA---P-RQDVWER-PHSE
                                          *
```

Fig 2. T-coffee multiple alignment of SH2-CD linker of ABL, TEC, SRC and CSK families.

Conclusions

The development of an InterProScan automated analysis has allowed us to study domain statistic distribution of kinase superfamily. This system is very versatile and will make a significant contribution in the demanding task of automatic classification approach of predicted proteins from genome sequencing projects.

Acknowledgements

MIUR “Functional Genomics” 449/97; FIRB “High throughput Grid computational platforms for virtual scalable organizations” and FIRB “Bioinformatics” projects and the centre of excellence recognized by the MURST C.I.S.I. (Centre for Biomolecular Interdisciplinary Studies and Industrial Applications).

References

[1] Hanks SK. Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. *Genome Biol*. 2003, 4(5):111.

[2] Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science*, 2002 Dec 6; 298(5600): 1912-34.

[3] Krupa A. and Srinivasan N. The repertoire of protein kinases encoded in the draft version of the human genome: atypical variations and uncommon domain combinations. *Genome Biol*. 2002, 3(12): 200.2.

[4] Gonfloni S. et al. The role of the linker between the SH2 domain and catalytic domain in the regulation and function of Src, *The EMBO J,* 1997, 16(24): 7261-71.