Cancer Immunity (28 December 2012) Vol. 12, p. 15

Cancer Immunity
Commentary

SurfaceomeDB: a cancer-orientated database for genes encoding cell surface proteins

Jorge Estefano Santana de Souza[1*], Pedro Alexandre Favoretto Galante[1*#], Renan Valieris Bueno de Almeida[1], Julia Pinheiro Chagas da Cunha[2], Daniel Takatori Ohara[1], Lucila Ohno-Machado[3], Lloyd J. Old[4] and Sandro José de Souza[1]

[1]*Ludwig Institute for Cancer Research, São Paulo Branch at Hospital Alemão Oswaldo Cruz, São Paulo SP, Brazil*

[2]*Center for Applied Toxicology, CAT/CEPID, Butantan Institute, São Paulo, Brazil*

[3]*Division of Biomedical Informatics, University of California, San Diego, San Diego, CA, USA*

[4]*Ludwig Institute for Cancer Research, New York Branch, New York, NY, USA*

[*]*These authors contributed equally to this work*
#Present Address: Hospital Sírio-Libanês, São Paulo, Brazil

Cell surface proteins (CSPs) are excellent targets for the development of diagnostic and therapeutic reagents, and it is estimated that 10-20% of all genes in the human genome encode CSPs. In an effort to integrate all data publicly available for genes encoding cell surface proteins, a database (SurfaceomeDB) was developed. SurfaceomeDB is a gene-centered portal containing different types of information, including annotation for gene expression, protein domains, somatic mutations in cancer, and protein-protein interactions for all human genes encoding CSPs. SurfaceomeDB was implemented as an integrative and relational database in a user-friendly web interface, where users can search for gene name, gene annotation, or keywords. There is also a streamlined graphical representation of all data provided and links to the most important data repositories and databases, such as NCBI, UCSC Genome Browser, and EBI.

Keywords: SurfaceomeDB, gene library, cell surface proteins

Introduction

The Human Genome Project and other related large-scale projects have provided an extraordinary amount of biomedical data to the public repositories. The organization of such data and the creation of user-friendly databases and webtools are important enterprises with a significant value to the whole research community. Presently, there are several databases/webtools publicly available and as an indication of their importance, many scientific journals have sections or issues dedicated exclusively to databases/webtools. A useful and updated list of database/webtools is available in the database issue of *Nucleic Acids Research* (1).

Among the most interesting sets of genes to be studied are those encoding cell surface proteins (CSPs). CSPs correspond to 10-20% of all coding genes in many eukaryote genomes (2) and are believed to act in many important cell functions as receptors, transporters, channels, and enzymes. Furthermore, they are excellent targets for diagnostic and therapeutic tools due to their subcellular localization. In a recent work (2), we explored the set of cell surface proteins (the surfaceome) in detail and realized how important it would be to have all information about CSPs organized in a database/webtool.

To address this issue, we developed SurfaceomeDB, a portal whose aim is to integrate a large variety of public information about the human surfaceome. SurfaceomeDB contains data related to gene annotation, gene expression, protein-protein interaction, and somatic mutations, among many other data types for all human genes encoding CSPs. A special emphasis is given to information related to human cancer. An efficient search system and a streamlined graphic representation allow the users to have an integrated view of several types of data, perform different queries, and retrieve useful information about any surfaceome gene.

Primary data

Seven major public data sources were used to build SurfaceomeDB: (i) transcript sequences from the Reference Sequences Project (3); (ii) gene annotation extracted from NCBI Gene Entrez (4); (iii) gene ontologies retrieved from the Gene Ontology Project (5); (iv) protein domains obtained from InterPro (6), PDB (7), and ModBase (8), including trans-membrane domains identified using TMHMM (9) and Pfam (10); (v) gene expression data obtained from SAGE Genie (11), MPSS database (12), and a large scale qPCR analysis (2) beside a link to the NCBI-GEO (13); (vi) protein-protein interactions obtained from a local database compiling data from public databases (14); (vii) somatic mutation data obtained from COSMIC database (15) and from a local compilation of all somatic mutations found in the literature; (viii) expression data from a variety of samples obtained from the Sequence Read Archive (SRA) maintained by the NCBI (16). All those data were processed and organized in a streamlined graphic representation in the SurfaceomeDB webtool. Table 1 summarizes all datasets used to build SurfaceomeDB, with the respective URLs for data retrieval. Further information can be obtained directly from the SurfaceomeDB web page. A dump of the respective MySQL database is also available for download.

Implementation

SurfaceomeDB runs on an Apache server with all pre-processed data stored in a MySQL 5.0 database. SurfaceomeDB web interface and graphical representations were built using CAKE-PHP. Data selection and data processing algorithms were built in PHP, Perl and shell scripts (see SurfaceomeDB website for more details about its implementation).

Table 1

Summary of all datasets used to build SurfaceomeDB.

Source	Link for Site
NCBI Gene Entrez	http://www.ncbi.nlm.nih.gov/gene/
CTD	http://ctd.mdibl.org/
Gene Ontology (GO)	http://www.geneontology.org/
Pathway (KEGG)	http://www.genome.jp/kegg/
TMHMM	http://www.cbs.dtu.dk/
InterPro	http://www.ebi.ac.uk/interpro/
Pfam	http://pfam.sanger.ac.uk/
PDB	http://www.pdb.org/
ModBase	http://modbase.compbio.ucsf.edu/
SAGE Genie	http://cgap.nci.nih.gov/SAGE/
MPSS	http://mpss.licr.org/
NCBI-GEO	http://www.ncbi.nlm.nih.gov/geo/
SRA (Sequence Read Archive)	http://trace.ncbi.nlm.nih.gov/sra/
qPCR	da Cunha *et al.* 2009
Protein-Protein Interaction	Cancherini *et al.* 2010
Somatic Mutations	COSMIC and a local database
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/

Web Interface

SurfaceomeDB is available at http://www.bioinformatics-Brazil.org/surfaceome. There is no use restriction and neither registration nor login is required. SurfaceomeDB web interface consists of a query section, a result summary, and a full result section (Figure 1).

The query section allows searches by gene symbol, gene symbol alias, NCBI Entrez Gene ID, and gene keywords. The search can also be done by chromosome regions and lists of gene names. Outputs are sorted by gene name, gene alias, and gene full annotation. Genes presenting the most similar gene names to the user's query are shown at the top of the 'Result Summary' section.

The result summary section shows the gene (official) name, gene (official) full name, and gene alias. This section allows users to quickly find if a gene is within the surfaceome set. For those belonging to the surfaceome set, full results are available by clicking the 'Gene Name' link.

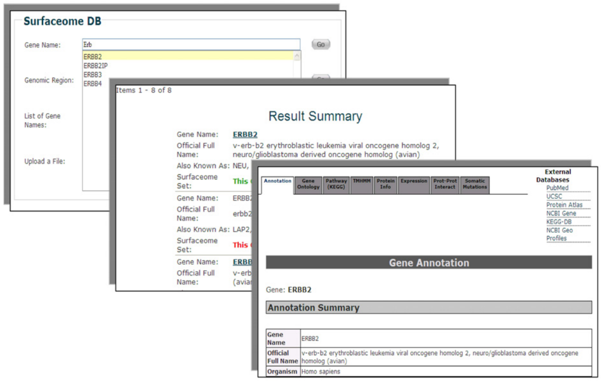
The 'Full Results' section is divided into 8 tabs containing different information. All tabs contain a menu at the top right corner containing links to external databases, such as PubMed (17), UCSC Genome Browser (18), Protein Atlas (19), NCBI Gene Entrez (4), and KEGG (20). The "Annotation" tab contains information about gene annotation, which includes gene name, gene alias, and a summary of gene function. In this tab, there is also an external link to the Comparative Toxicogenomics Database (CTD) (21), which presents information about chemical molecules that interact with the respective cell surface protein.

The "Gene Ontology" tab presents Gene Ontology (GO) classification for the selected gene. That includes information about Biological Process, Molecular Function, and Cellular Component, as classified by GO. All GO classifications have an external link to AmiGO, the official web-based set tools for searching and browsing the Gene Ontology database. The "KEGG Pathway" tab contains information about known signaling pathways in which the surfaceome genes are involved. All signaling pathways are based on KEGG's data. Results in this tab section contains KEGG pathway ID, pathway name, and an external link to the KEGG website.

The "TMHMM" tab reports the output of the TMHHM program (9) with information about transmembrane domains found in the respective protein. Transmembrane domains located in the first 50 amino acids (amino terminal region) were considered as signal peptides and were excluded from the surfaceome set. The "Protein Info" tab reports a series of information for the respective protein, including domain composition, as well as 3D structure, when available. The "Expression" tab contains gene expression information. Data from three gene expression technologies, SAGE Genie (11)/ MPSS (12), qPCR, and microarrays, were used to infer gene expression. Short and Long SAGE data were retrieved from SAGE Genie (11). qPCR data were obtained from da Cunha *et al.* (2) and organized in a graphical representation. Microarray data were downloaded from NCBI-GEO (13) and only those studies using human samples were selected and organized in a graphical representation.

The "Protein-Protein Interaction" tab contains a graphical representation of PPI data obtained from a local database (14), compiled from several PPI datasets. Finally, the "Somatic Mutation" tab reports somatic mutations identified for the respective gene in a variety of tumor types. Data for this tab was compiled from different sources, including COSMIC (15) and reports from the literature. For each gene there is a graphic representation showing all respective mutations indexed according to genomic, cDNA, and protein coordinates. Additional information about the somatic mutation(s), such as the tumor tissue where the mutation is found, the mutation type (synonymous or non-synonymous), and genomic position are also provided.

Figure 1



Surfaceome DB

Gene Name: Erb

ERBB2
ERBB2IP
ERBB3
ERBB4

Items 1 - 8 of 8

Result Summary

Gene Name: ERBB2
Official Full Name: v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
Also Known As: NEU,
Surfaceome Set: This
Gene Name: ERBB2
Official Full Name: erbb2
Also Known As: LAP2,
Surfaceome Set: This
Gene Name: ERBB
Official Full Name: v-erb (avian

Gene Annotation

Gene: ERBB2

Annotation Summary

Gene Name	ERBB2
Official Full Name	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
Organism	Homo sapiens

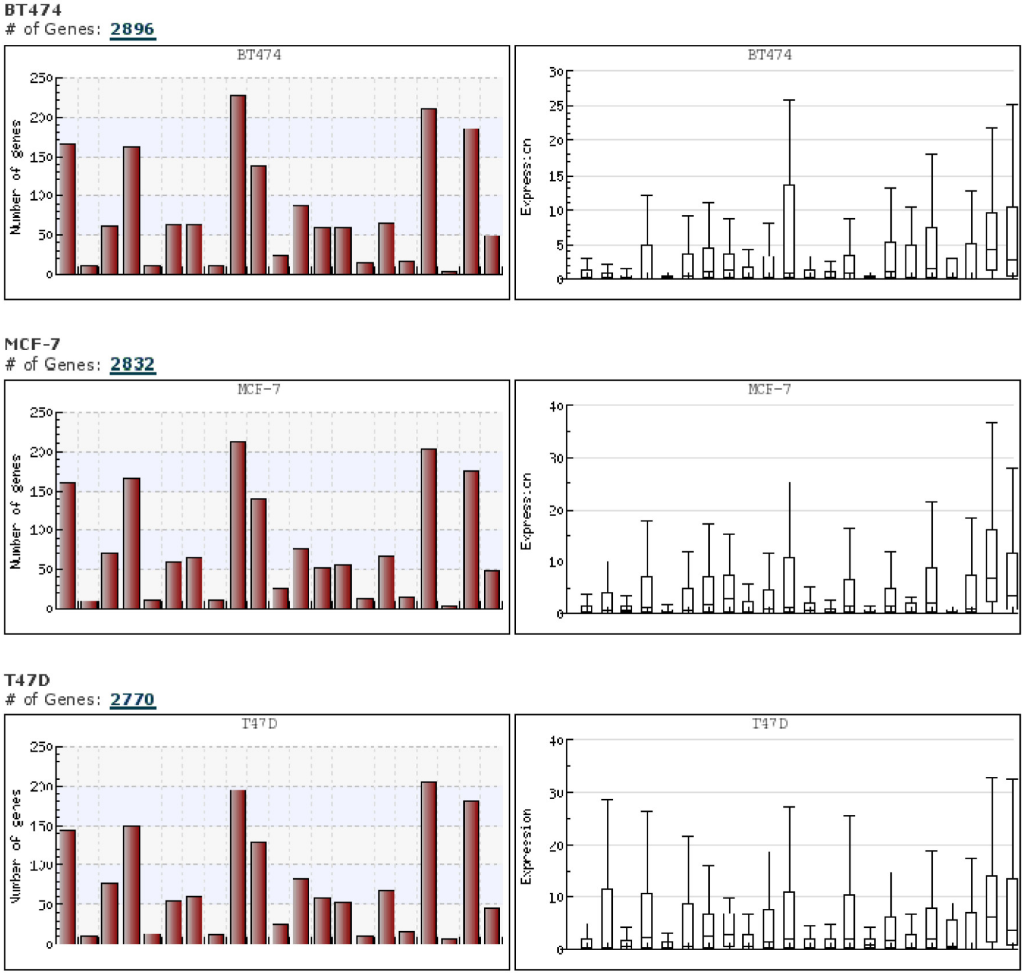
SurfaceomeDB web portal is divided in three parts: a query section, a results summary section, and a full results section. This last section contains a variety of data provided in a gene-centered fashion.

Surfaceome Display

It is reasonable to envisage that in the next few years a large amount of gene expression data will be available for a large variety of biological samples, including normal/tumor cell lines, and even single-cell preparations. This is due to the significant impact of next-generation sequencing technologies in gene expression analysis. This deeper coverage of a given transcriptome allows, in principle, an exhaustive profiling of all genes expressed in that cell/tissue. The use of tumor cell lines, for example, would allow the unambiguous identification of genes exclusively/differentially expressed due to the absence of normal cells.

To make use of such data, an application (called Surfaceome Display) was implemented in the SurfaceomeDB to profile the expression pattern of one or more libraries and to compare the

Figure 2



Expression profiling of the surfaceome for the three breast cancer cell lines BT474, MCF-7, and T47D. Charts in the left report the number of genes expressed in the respective cell line within each surfaceome category. Boxplots in the right report the expression level of genes within each surfaceome category.

expression profiling of two groups of libraries. The expression profiling of any given library provides a series of information including all genes expressed and a categorization of expressed genes based on cell surface protein families. Figure 2 illustrates a comparison of the expression profiles of three breast tumor cell lines (BT474, MCF-7, and T47D).

By using Surfaceome Display, users can also identify genes exclusively expressed in one group of samples or genes differentially expressed between the two groups. A series of graphs and annotations are provided in the 'Results' section. To illustrate this use of the Surfaceome Display, genes expressed in three breast tumor cell lines (BT474, MCF-7, and T47D) were compared to genes expressed in the normal breast cell line HME plus a panel of ten normal tissues (cerebellar cortex, adipose, brain, breast, colon, heart, liver, lymph node, skeletal muscle, and testis).

First, all genes expressed in the normal panel were used to filter genes expressed in the three breast tumor cell lines. Eight genes were found to be exclusively expressed in the breast tumor cell lines (GPR139, OR1J1, OR1J4, OR1L6, OR1N1, OR1N2, OR2G2, and TAS2R43). The high proportion of olfactory receptors (6 out of 8) is probably due to the very restricted expression pattern of these genes in normal tissues. Their expression in tumors, however, raises the possibility that they can be used as targets for therapeutic intervention, as suggested by others (22).

Table 2

Genes upregulated in the breast tumor cell lines.

Genes							
ABCA4	CALCR	DSCAM	GRPR	LRRC59	PPAP2C	SLITRK6	TMEM45B
AMIGO2	CD207	EMP2	HFE	LRRC8E	PRLR	SPINT1	TMEM49
ANKRD22	CD24	ERBB2	IFI6	MARVELD2	PRRG2	SPINT2	TMEM64
ATP2C2	CDH1	ERBB3	IGF1R	MARVELD3	PVRL4	ST14	TMPRSS13
ATP6V0A4	CEACAM16	FIBCD1	IL1RAPL2	MC3R	RAG1AP1	STARD3	TNFRSF12A
ATP8B1	CELSR1	FOLR1	ILDR1	MCOLN2	SHISA2	SVOPL	TNFSF15
BAMBI	CHRNA5	FRAS1	KCNK6	MUC1	SIGLEC15	TARP	TPBG
BIK	CLDN3	FREM2	KIAA1324	NAALADL2	SLC12A8	TFRC	TRPV6
BRI3BP	CLDN4	GJD3	KLRG2	NCAM2	SLC2A10	TLCD1	UPK1A
C18ORF45	CLDN7	GPR132	KREMEN2	NPY1R	SLC34A3	TM7SF2	UPK2
C1ORF210	COLEC12	GPR143	LAPTM4B	OCLN	SLC39A6	TMC4	XKRX
C3ORF57	CRB3	GPR160	LCT	OR1Q1	SLC5A8	TMEM125	ZDHHC12
CA12	CYB561	GPR172B	LDLRAD1	P2RX2	SLC6A14	TMEM139	ZP3
CACNG4	DEGS2	GPR77	LPCAT1	P2RY2	SLC6A3	TMEM194A	
CACNG5	DNAJC22	GPR81	LRRC26	P2RY6	SLC9A2	TMEM30B	

To identify genes upregulated in the three breast tumor cell lines, when compared to all normal samples, we set a threshold of fivefold difference. That gave us a total of 118 genes upregulated in the breast tumor cell lines, when compared to all normal tissues (Table 2).

Table 3

Genes exclusively expressed in the ductal carcinoma cell lines.

Genes							
ADAM2	CCR9	GHSR	MC3R	OR1B1	OR4D1	SORCS3	TSPAN16
ADAM30	CDH7	GPR151	MDGA2	OR1J1	OR51B4	SPEM1	UMODL1
AQP2	CHRM2	IMPG1	MEP1B	OR1J4	OR5C1	TARP	VN1R5
AREG	CHRNA7	KCNH7	MS4A3	OR1L6	OR8S1	TAS2R42	
ARSH	CLRN1	LDLRAD1	NTSR2	OR1N1	SCN1A	TAS2R43	
BTLA	DCC	LECT1	ODF4	OR1N2	SLC10A5	TRHR	
CACNG5	DPP10	LHFPL1	OPRD1	OR1Q1	SLC12A1	TRPM5	
CCR6	GABRA5	LRTM2	OR10J1	OR2G2	SLC7A13	TRPM8	

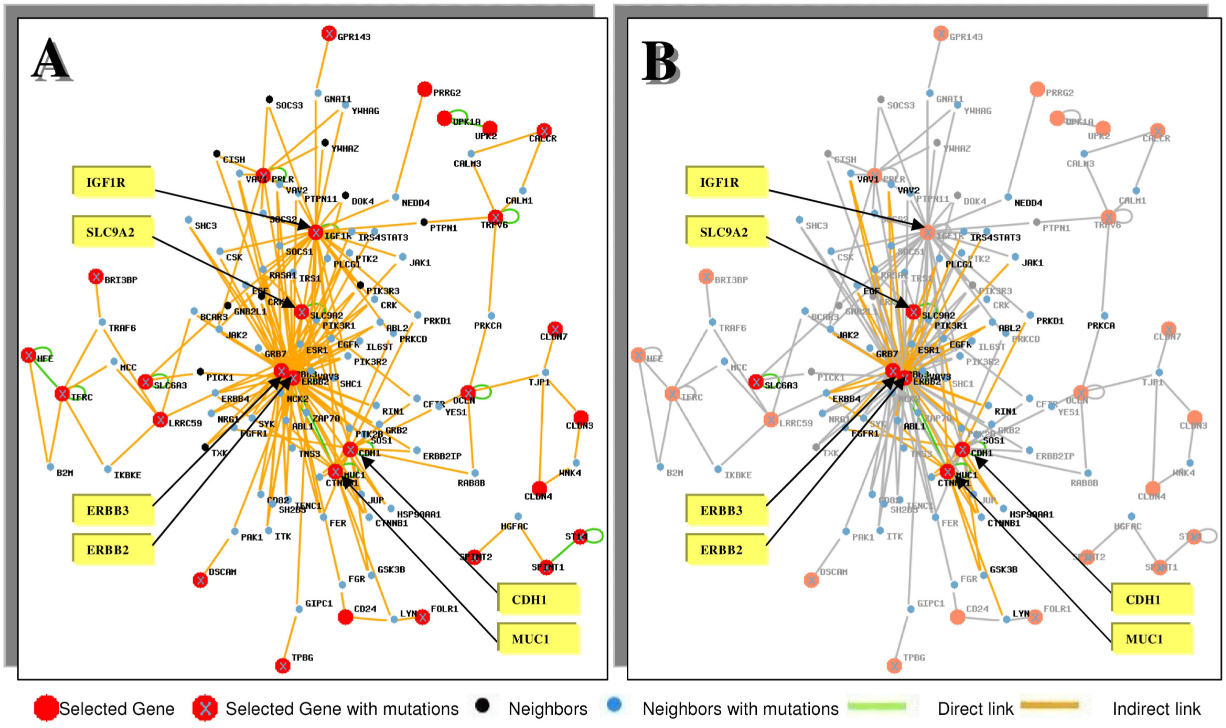
The computational strategy used in the Surfaceome Display allows even comparisons between the breast tumor cell lines. BT474 and T47D, for example, are ductal carcinomas while MCF-7 is an adenocarcinoma. Comparisons between these cells could provide candidates for markers of each tumor type. There are, for example, 59 surfaceome genes (Table 3) that are exclusively expressed in the ductal carcinoma cell lines, when compared to MCF-7, the normal cell line HME, and normal breast.

An interesting feature of the Surfaceome Display is the possibility of integrating different types of information into a PPI network scaffold. Figure 3 illustrates this feature. All 118 genes upregulated in the breast tumor cell lines were integrated into a PPI network scaffold. It is possible to visualize that most genes are centralized around six major hubs: ERBB2, ERBB3, IGF1R, CDH1, MUC1, and SLC9A2 (Figure 3A). When somatic mutations found in breast tumors are integrated into the same network, we observe that five out of the six hubs described above are mutated in breast cancer (Figure 3B). Most of the interacting partners of these five hubs are also mutated in breast cancer. This type of integrative view of cancer-related data opens new opportunities for the development of more effective therapeutic and diagnosis protocols.

Discussion

We make available to the community SurfaceomeDB, a database integrating information on human CSPs with a special emphasis on cancer-related data. With the impressive development of sequencing technologies, we envisage that the amount of genetic information will continue to increase at an exponential rate. Databases restricted to a certain subset of genes/proteins are important in the sense that they provide more specific information and associations that would otherwise be absent (or diluted) in genome-scale databases. We are confident that SurfaceomeDB will be a helpful resource to the community.

Figure 3



PPI network for genes differentially expressed in breast tumor cell lines. (A) All 118 genes upregulated in the breast tumor cell lines were integrated into a PPI network scaffold. (B) Somatic mutations found in breast tumors are integrated into the same network of 118 genes upregulated in the breast tumor. We observe that five out of the six hubs described above are mutated in breast cancer.

Abbreviations

CSPs, Cell Surface Proteins

Acknowledgements

The authors thank Patricia M. de Carvalho for technical support. Part of this work was supported by grants D43TW007015 from the Fogarty International Center, National Institutes of Health, and 2007/55790-5 from Fundação de Amparo à Pesquisa do Estado de São Paulo.

References

1. Galperin MY, Cochrane GR. The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res* 2011; **39:** D7-D10. (PMID: 21177655)

2. da Cunha JP, Galante PA, de Souza JE, de Souza RF, Carvalho PM, Ohara DT, Moura RP, Oba-Shinja SM, Marie SK, Silva WA Jr, Perez RO, Stransky B, Pieprzyk M, Moore J, Caballero O, Gama-Rodrigues J, Habr-Gama A, Kuo WP, Simpson AJ, Camargo AA, Old LJ, de Souza SJ. Bioinformatics construction of the human surfaceome. *Proc Natl Acad Sci USA* 2009; **106:** 16752-16757. (PMID: 19805368)

3. *RefSeq: Reference Sequences Project.* Accessed from: http://www.ncbi.nlm.nih.gov/RefSeq/

4. *NCBI Gene Entrez.* Accessed from: http://www.ncbi.nlm.nih.gov/gene/

5. *GO: Gene Ontology Project.* Accessed from: http: www.geneontology.org/

6. *InterPro: Integrated Database of Predictive Protein Signatures.* Accessed from: http://www.ebi.ac.uk/interpro/

7. *PDB: Protein Data Bank.* Accessed from: http://www.pdb.org/

8. *ModBase: Database of Comparative Protein Structure Models.* Accessed from: http://modbase.compbio.ucsf.edu/

9. *TMHMM: TransMembrane Prediction Using Hidden Markov Models.* Accessed from: http://www.cbs.dtu.dk/services/TMHMM/

10. *Pfam: Protein Families.* Accessed from: http://pfam.sanger.ac.uk/

11. *SAGE Genie.* Accessed from: http://cgap.nci.nih.gov/SAGE/

12. *MPSS Database.* Accessed from: http://mpss.licr.org/

13. *NCBI-GEO: Gene Expression Omnibus.* Accessed from: http://www.ncbi.nlm.nih.gov/geo/

14. Cancherini DV, França GS, de Souza SJ. The role of exon shuffling in shaping protein-protein interaction networks. *BMC Genomics* 2010; **11 Supplem. 5:** S11. (PMID: 21210967)

15. *COSMIC: Catalogue of Somatic Mutations in Cancer.* Accessed from: http://www.sanger.ac.uk/genetics/CGP/cosmic/

16. *SRA: Sequence Read Archive.* Accessed from: http://www.ncbi.nlm.nih.gov/SRA/

17. *PubMed.* Accessed from: http://www.ncbi.nlm.nih.gov/PubMed/

18. *UCSC Genome Browser.* Accessed from: http://genome.ucsc.edu/

19. *The Human Protein Atlas Project.* Accessed from: http://www.proteinatlas.org/

20. *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Accessed from: http://www.genome.jp/kegg/

21. *CTD: Comparative Toxicogenomics Database.* Accessed from: http://ctd.mdibl.org/

22. Neuhaus EM, Zhang W, Gelis L, Deng Y, Noldus J, Hatt H. Activation of an olfactory receptor inhibits proliferation of prostate cancer cells. *J Biol Chem* 2009; **284:** 16218-16225. (PMID: 19389702)

Contact

Address correspondence to:

Sandro José de Souza, Ph.D.
Ludwig Institute for Cancer Research
São Paulo Branch at Hospital Alemão Oswaldo Cruz
Laboratory of Computational Biology
João Julião St, 245, 1st floor
01323-903 São Paulo
Brazil
Tel.: + 55 11 3388 3248
Fax: + 55 11 3141 1325
E-mail: sandro@compbio.ludwig.org.br