IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

INTEGRATIVE BIOINFORMATICS ANALYSIS FOR EXTRACTING INSIGHTFUL PATTERNS FROM GENETIC SEQUENCES

Rajendra Soni[1], Dr. Amrita Verma[2]
[1]Research Scholar, Assistant Professor
[1]Department of Computer Science Engineering
[1]Dr. C.V. Raman University, Kota, Bilaspur, India

Abstract: The widespread availability of nucleotide and amino acid sequence data has enabled the development of advanced methods for identifying biologically and clinically significant information. Online platforms such as GeneCards offer accessible repositories for extended research and academic exploration. Additional protein-related data can be retrieved from databases like UniProt and SwissProt. Tools such as FASTA and Clustal_X are commonly used to calculate sequence similarity, aiding in the selection of target genes. In addition, text mining serves as a valuable technique for extracting relevant knowledge from scientific literature.

Index Terms - **FASTA, Clustal_X, Genomic Data, Genetic Sequences, Bioinformatics**

I. INTRODUCTION

Genetic sequence data, including nucleotide and amino acid sequences from a variety of organisms such as plants, bacteria, and animals, is now widely available. The current focus has moved toward the efficient retrieval, annotation, and meaningful application of this existing information[1]. Collaborative efforts among various interdisciplinary scientific communities are playing a vital role in advancing genetic research. This article presents a concise summary of the major bioinformatics resources associated with nucleotide and protein sequence data.

II. GENE CARDS

In the 1990s, the World Wide Web emerged as a major platform for hosting vast amounts of biological information. Despite its richness, the overwhelming volume and complex web of hyperlinks often made it difficult for users to access relevant data efficiently[2-3]. To address this challenge, GeneCards was created as a resource designed to present gene-related information in a structured and conceptually coherent format. It offered detailed insights into human genes, including their functions and associated medical conditions.

The data is organized in a flat file structure and indexed using the Glances tool. Gene-specific information is generated through a CGI script, offering users straightforward and efficient access. These resources are widely available for scientific use and act as a unified platform for integrating human genetic data related to genes, proteins, and diseases[4]. Over time, it has evolved into a key resource for compiling and extracting meaningful biological information.

III. HOW TO USE

To start a search on the GeneCards platform, type your query into the "Search GeneCards" input field and either press the Enter key or click the button situated to its right [5].

Example of a keyword-based search: diabetes AND tongue

The search results page presents a brief list of minicards, each displaying key details such as the gene symbol, description, category, GCID, and a relevancy score. To view more information within a minicard, click the plus (+) icon to the left of the gene symbol. The expanded GeneCard will highlight all sections where your search term(s) appear. Additionally, the minicards will highlight all keywords entered in the search, including any stemmed word variations [6].

Select the gene you wish to explore further by clicking on its gene symbol, which appears on the left side under the 'Symbol' column.

The GeneCard provides detailed and extensive information about the selected gene.

Gene statistics are located at the bottom of the GeneCards webpage. The first column shows the total number of genes, along with the count of those officially approved by the HGNC (HUGO Gene Nomenclature Committee). These figures are linked to graphs illustrating gene distribution [7]. Additionally, this column includes hyperlinks to some of the most notable GeneCards entries, featuring both highly referenced genes and those linked to specific diseases.

The second and third columns display the number of genes in each category, along with links to a statistics page. This page offers gene distribution charts and tools for searching genes within specific categories.

The last column provides links to representative genes for each category.
At the bottom of each page, there are links to the GeneCards gene index, offering a complete list of all genes included in the GeneCards database.

Table 1. GeneCard

GeneCards Version 5.19 (Updated: Jan 24, 2024)

Total genes	4,66,053	Category	# of Genes	Example Genes
HGNC approved	43,809	Protein-coding	21,617	MTOR FGFR2 RET MET MAP2K1 IGF1R FGFR3
Disease genes	20,197	ncRNA genes	2,91,142	
Hot genes	500	lncRNAs	1,30,000	SFTA3 OFCC1 HCP5 C10orf55 SLC22A18AS DLEU1 WT1-AS
		piRNAs	1,11,811	piR-52356 piR-34881 piR-30791-073 piR-62069 piR-62060 piR-62024 piR-61955
		miRNAs	6,903	MIR21 MIR27A MIR145 MIR143 MIR140 MIRLET7D MIRLET7A1
		rRNAs	1,250	MT-RNR2 MT-RNR1 RNA5S17 RNA5S16 RNA5S15 RNA5S13 RNA5S12
		tRNAs	1,158	MT-TL1 MT-TT MT-TS1 MT-TW MT-TV MT-TL2 MT-TK
		snoRNAs	1,904	SNORD3A SNORD24 SNORD118 SNORA75 SNORA73B SNORA66 SNORA64
		SRP RNAs	9,022	RN7SL1 RN7SL2 RN7SL3 RF00017-7992 RF00017-7752 RF00017-6963 RF00017-6018
		circRNAs	120	OP794511 OP794616 OP794610 OP794600 OP794560 OP794534 OP794524
		Other ncRNAs	28,974	TERC ADGRF2P ARRDC1-AS1 RNU4ATAC HCG22 SCARNA6 SCARNA8
		Functional elements	1,28,259	FRAXA FRAXE HBB-LCR LOC109504725 LOC107982234 LOC107303338 LOC107133510
		Pseudogenes	21,932	GGT2P CASTOR3P SLC26A10P GUCY1B2 NXF5 BIRC8 TRIM16L
		Genetic loci	1,288	ERVE-1 ST2 PCAP IGES AZF1 DCR PARK16
		Gene clusters	10	IGKV@ PCDHB@ PCDHG@ HOXA@ HOXD@ IGLV@ IFN1@
		Uncategorized	1,805	UGT1A C20orf181 AKAP2 ENSG00000262202 PALM2 CCDST ENSG00000266919

IV. UNIPROT AND SWISSPROT

UniProt is a comprehensive resource dedicated to protein-related information. Its primary goal is to provide a well-structured, extensively annotated, and accurate database of protein sequences, supported by numerous cross-references and user-friendly interfaces [8]. The UniProt Archive (UniParc) stores all publicly available protein sequence data, while the UniProt Knowledgebase (UniProtKB) functions as the central repository, offering consistent and detailed functional annotations. Additionally, the UniProt NREF database, derived from UniProtKB, delivers a non-redundant dataset that ensures broad coverage of protein sequence space at various levels of resolution. This initiative aims to support researchers and scholars by integrating large-scale data from the Human Genome Project, as well as from structural genomics, functional genomics, and proteomics studies [9].
The SwissProt protein knowledgebase connects amino acid sequences with up-to-date information across various areas of the life sciences. Its strength lies in its high-quality annotations, use of standardized terminology, direct links to specialized databases, and minimal redundancy. SwissProt is deeply integrated with a wide range of expert resources, enabling users to efficiently navigate and explore the current scientific understanding of proteins. This level of integration offers significant insights into the complex and dynamic world of protein biology [10].

>sp|O95905|ECD_HUMAN Protein ecdysoneless homolog OS=Homo sapiens OX=9606 GN=ECD PE=1 SV=1

MEETMKLATMEDTVEYCLFLIPDESRDSDKHKEILQKYIERIITRFAPMLVPYIWQNQPF

NLKYKPGKGGVPAHMFGVTKFGDNIEDEWFIVYVIKQITKEFPELVARIEDNDGEFLLIE

AADFLPKWLDPENSTNRVFFCHGELCIIPAPRKSGAESWLPTTPPTIPQALNIITAHSEK

ILASESIRAAVNRRIRGYPEKIQASLHRAHCFLPAGIVAVLKQRPRLVAAAVQAFYLRDP

IDLRACRVFKTFLPETRIMTSVTFTKCLYAQLVQQRFVPDRRSGYRLPPPSDPQYRAHEL

GMKLAHGFEILCSKCSPHFSDCKKSLVTASPLWASFLESLKKNDYFKGLIEGSAQYRERL

EMAENYFQLSVDWPESSLAMSPGEEILTLLQTIPFDIEDLKKEAANLPPEDDDQWLDLSP

DQLDQLLQEAVGKKESESVSKEEKEQNYDLTEVSESMKAFISKVSTHKGAELPREPSEAP

ITFDADSFLNYFDKILGPRPNESDSDDLDDEDFECLDSDDDLDFETHEPGEEASLKGTLD

NLKSYMAQMDQELAHTCISKSFTTRNQVEPVSQTTDNNSDEEDSGTGESVMAPVDVDLNL

VSNILESYSSQAGLAGPASNLLQSMGVQLPDNTDHRPTSKPTKN

Figure 1. Extract of Protein Sequence

V. FASTA (SEQUENCE COMPARISON)

Once a specific amino acid or nucleotide sequence is selected from a database, software tools like FASTA can be used to perform sequence comparisons. This program evaluates similarity scores and identifies structural relationships by analyzing sequence similarity. In the initial comparison phase, it detects regions where sequences share similarities. These regions are then reassessed using a scoring matrix that also considers shorter sequence matches, contributing to the overall similarity score [11]. The algorithm then focuses on the highest-scoring starting region to establish an optimal alignment between initial segments. Finally, sequences with the top similarity scores are aligned using a refined optimization strategy [12].

By applying bioinformatics techniques, researchers can select specific sequences and analyze the data to interpret gene functions, disease-related expressions, and biological pathways. Emerging methods are being developed to automate this process. The main goal is to identify references to relevant biological entities—such as genes, proteins, and related components—within textual data and generate automated annotations for these proteins. Tools like PathBinder and GENIA utilize natural language processing (NLP) to facilitate text mining in biological research.

REFERENCES

[1] C. Tanford, J.A. Reynolds, 2001. Nature's Robots: A History of Proteins, Oxford University Press, Oxford, New York.

[2] E.D. Levy, E. Boeri Erba, C.V. Robinson, S.A. Teichmann, 2008. Assembly reflects evolution of protein complexes, Nature 453 1262–1265.

[3] F. Alber, S. Dokudovskaya, L.M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, A. Sali, M.P. Rout, 2007. The molecular architecture of the nuclear pore complex, Nature 450, 695–701.

[4] A.C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L.J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M.A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J.M. Rick, B. Kuster, P. Bork, R.B. Russell, G. 2006. Superti-Furga, Proteome survey reveals modularity of the yeast cell machinery, Nature 440, 631–636.

[5] M. Chance, 2008. Mass Spectrometry Analysis for Protein–Protein Interactions and Dynamics, Wiley, Hoboken, NJ.

[6] A. Brückner, C. Polge, N. Lentze, D. Auerbach, U. Schlattner, 2009. Yeast Two-Hybrid: a powerful tool for systems biology, Int. J. Mol. Sci. 10, 2763–2788.

[7] N. Ramachandran, J.V. Raphael, E. Hainsworth, G. Demirkan, M.G. Fuentes, A. Rolfs, Y. Hu, J. LaBaer,2008. Next-generation high-density self-assembling functional protein arrays, Nat. Meth. 5, 535–538.

[8] J. Mulder, E. Bjorling, K. Jonasson, H. Wernerus, S. Hober, T. Hokfelt, M. Uhlen, 2009.Tissue profiling of the mammalian central nervous system using human antibody-based proteomics, Mol. Cell Proteomics 8, 1612–1622.

[9] J. Fasolo, M. Snyder,2009. Protein microarrays, Meth. Mol. Biol. 548, 209–222.

[10] T.H. Patwa, Y. Qiu, J. Zhao, D.M. Simeone, D.M. Lubman, 2009. All-liquid separations, protein microarrays, and mass spectrometry to interrogate serum proteomes: an application to serum glycol proteomics, Meth. Mol. Biol. 520, 75–87.

[11] J Hu, X. He, K.A. Baggerly, K.R. Coombes, B.T. Hennessy, G.B. Mills, 2007. Non- parametric quantification of protein lysate arrays, Bioinformatics 23, 1986–1994.

[12]Y. Gu, A. Zeleniuch-Jacquotte, F. Linkov, K.L. Koenig, M. Liu, L. Velikokhatnaya, R.E. Shore, A. Marrangoni, P. Toniolo, A.E. Lokshin, A.A. Arslan, 2009. Reproducibility of serum cytokines and growth factors, Cytokine 45, 44–49.