VOLUME - 13, ISSUE - 09, SEPTEMBER - 2024 • PRINT ISSN No. 2277 - 8160 • DOI : 10.36106/gjra

Original Research Paper

Bioinformatics

NEQUIM CONTACT SYSTEM (NCS): A TOOL FOR THE GENERATION AND ANALYSIS OF PROTEIN-LIGAND INTERACTION FINGERPRINTS

MS. Andrelly Martins-Jose — Chemoinformatics Group - Nequim, Departamento De Quimica, Icex, Ufmg, Brazil.

Dr. Vera Lucia de Almeida — Serviço De Fitoquímica E Prospecção Farmacêutica, Fundação Ezequiel Dias, Belo Horizonte, Mg, Brazil.

Prof. Dr. Julio Cesar Dias Lopes* — Chemoinformatics Group - Nequim, Departamento De Quimica, Icex, Ufmg, Brazil. *Corresponding Author

ABSTRACT NEQUIM Contact System (NCS) is a system that generates and analyzes interaction vectors of protein-ligand complexes. Core features include multiple views of vectors, multiple selection options, cluster analysis, and the generation of interaction vector models. The input could be from a PDB format or files generated by automatic docking software AutoDock, Vina, or Surflex. Availability: The NCS is available free of charge from the SourceForge website https://sourceforge.net/projects/nequimcontacts

KEYWORDS : protein-ligand complex, automatic docking, intermolecular interaction; interatomic contact, Linux software, bioinformatics, chemoinformatics

INTRODUCTION

Biological processes rely on intermolecular interactions between biomolecules and their receptors, which are crucial pharmacological investigations of diseases. These interactions can be studied through experimental methods or by analyzing three-dimensional structures obtained via X-ray or NMR techniques. Various software tools exist for analyzing protein-ligand complexes, such as HBPlus (McDonald et al., 1994) and LPC (Sobolev et al., 1999), but they often lack user-friendly interfaces. Tools like FingerPrintLib (Desaphy et al., 2013) and SIFt (Deng et al., 2004) utilize interaction fingerprints for analysis, but SIFt is not widely distributed, and FingerPrintLib functions as a plugin for PyMol.

The NEQUIM Contact System (NCS) was developed to efficiently process extensive datasets from docking programs. It identifies intermolecular contacts and generates interaction fingerprints that can be manipulated and compared through an interactive interface.

Implementation

NCS is a stand-alone application implemented in FreePascal and distributed for the Linux platform. It provides a graphical interface for analyzing intermolecular interactions among ligands and their biological targets. NCS uses LPC as a contact engine program that calculates the intermolecular contacts from standard three-dimensional structure PDB files. The external programs ClustalW (Larkin et al., 2001) and Cluster2.9 are used for sequence alignment and cluster analysis, respectively.

Input and Output Data

NCS works with three-dimensional structures of protein-ligand complexes in PDB format or from output files generated by AutoDock4 (Morris et al., 2009), Vina (Eberhardt et al., 2021) or Surflex (Jain et al., 2003) docking programs. NCS can also upload data produced by AutoDock, Vina, or Surflex docking programs. The complete analysis can be stored and retrieved in NCS standard format (ncs extension) or plotted in PNG graphical format. Additionally, the vectors can be imported or exported in bitstring format. The models produced by analyzing multiple complexes (see Contact Models section), which can be used for similarity analyses and ranking of different complexes, are stored in NCS Model format (mncs extension).

Interaction vectors

The vectors can be handled quickly, even in large quantities, and submitted to clustering or other analyses, such as score analysis. The intermolecular contacts in the ligand-receptor complex are computed using the LPC program and translated into a binary vector where the interactions of each amino acid residue of the sequence are encoded in a six-bit long string. The six bits correspond to i. existence of contact; ii. contact with side chain; iii. hydrogen bonding contact; iv. hydrophobic contact; v. aromatic contact; and vi. destabilizing contact. The classification of contacts is based on Sobolev (Sobolev et al., 1999) atom types (classes) and the distance between interacting atoms.

Sequence alignment

To compare the binding mode of ligands with different proteins, it is necessary to make the previous alignment of the sequences. The goal of aligning sequences is to identify the best correlation among amino acid residues of each sequence. The sequence alignment allows a more accurate comparison among the fingerprints as they will be compared to equivalent positions of the sequences. NCS calls the external program ClustalW to make the multiple sequence alignment. After aligning sequences, NCS automatically realigns the interaction vectors, filling the gaps with zeros.

Similarity calculations

After the sequence alignment, NCS compares the entire vector set (all-against-all), and the similarity of each pair of vectors is calculated using the Tanimoto coefficient Equation (1). The Tanimoto index is calculated by considering the size of the vectors A (N_A) and B (N_B), i.e., the sum of the bits turned on for each vector and the sum of the bits in common to vectors A and B (N_{AB}):

$$TC = \frac{N_{AB}}{N_A + N_B - N_{AB}} \quad (1)$$

The Frequency Model (see the section Interaction vector Models) is a vector defined by Equation (2), where F1, F2, …, F_N are the vectors 1, 2, …, N, and N is the number of vectors analyzed. The model is essential for analyzing individual interaction vectors and evaluating their similarity to the entire dataset.

$$FM(i) = \frac{F_1(i) + F_2(i) + \cdots + F_N(i)}{N} \quad (2)$$

When comparing the interaction vector of an individual complex with a model generated from multiple vectors, we also implemented a Score based on the Frequency Model (SFM), defined by Equation (3). This score quantitatively

measures the similarity between the individual complex and the model. In this equation, bit(i) is the value of the bit ith of the complex, and FM(i) is the frequency of bit i in the Frequency Model. The SFM score can be used to rank and compare different complexes based on their similarity to the model:

$$SFM = \sum bit(i) * FM(i) \qquad (3)$$

Clustering

After generating the similarity matrix, the interaction vectors are clustered by the program Cluster2.9. This cluster analysis is done automatically by NCS and can be easily redone after manipulating (adding or removing) the vector set under study. The clustering results generate a phylogenetic tree, presented at the graphical interface along with the aligned sequences and the vectors.

Interface

The graphical interface of NCS provides options for creating, manipulating, and analyzing interaction vectors (Figure 1). The Menu has options to read or save the analyses (Menu File), to make selections (Menu Select), to change the view (Menu View), and to perform calculations (Menu Run). The selection options allow the user to select up to five groups. These groups can be manipulated separately or highlighted with different colors. The options include adding or removing individual complexes, generating a model from multiple selections of complexes, or saving them as a new project. The view options include the choice of background color for each selected group, coloring the bits according to interaction type, and the alignment sequence color. These colored bits of the interaction permit specific interactions to be highlighted and help understand the importance of these interactions in particular systems. There are three alignment sequence color modes: highlight the residues that contact the ligand, highlight the muted residues, and the ClustalW scheme.

Interaction vector Models

Most bioactive compounds trigger a specific response and should have specific contact with their biological targets. The idea behind the structure-based virtual screening is that it would be possible to discriminate between active and inactive compounds by carefully analyzing the contacts between the molecule and its biological target.

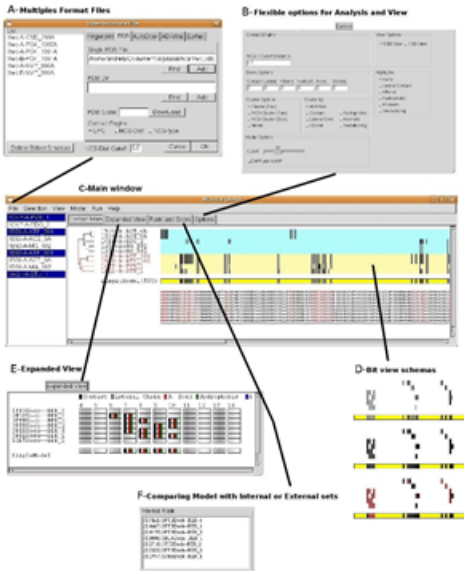

Figure 1 – NEQUIM Contact System (NCS) Interface. The NCS provides several options for selection, coloring, calculations, manipulations, and display to facilitate the analysis of a large set of interaction vectors.

From a set of three-dimensional structures that include some active compounds bound to the same or similar proteins, NCS can represent the contacts between the entire set of ligands and receptor(s). This model can be used as a reference for filtering structures obtained by docking or to better understand the interactions associated with recognizing and binding a specific ligand with its biological receptor. The models generated by the NCS are based on the frequency of bits within the vectors utilized for model creation. This model illustrates the occurrence rate of specific interactions within the chosen set of three-dimensional structures.

Analyzing a specific complex vector against a model can determine if a ligand's binding mode aligns with established modes for that protein. Different ligand binding modes are linked to specific receptor interactions, as noted by Bender et al. (2004). This analysis allows for the ranking of docking structures based on their binding modes besides docking energy filtering. Additionally, this method is an alternative for selecting ligands with various binding modes during post-docking evaluations.

The score and posterior ranking of the complex vector set are done in two different ways. The first uses a fixed cutoff, forming a binary vector model (BFModel) that considers only the bits with an equal or higher frequency cutoff. Alternatively, each vector position is weighted by its frequency in the model. The score is the sum of the multiplication of each bit by the frequency of the same bit in the Frequency Model (see Similarity Section) (Xue et al., 2003).

Post-processing of Docking Simulations

The biological activity of small molecules results from specific interactions with their targets, which can be quantified using interaction vectors generated by the NCS program. Pursuing higher success rates has led to improved docking techniques and post-docking analyses to minimize the number of false positives in the selection of hits. Interaction models derived from experimental data serve as filters or refinement parameters in docking, allowing for the assignment of weights to ligand poses based on descriptors that differentiate active from inactive compounds. The interaction vectors obtained through docking can be compared with those from crystallography or previously constructed models, facilitating the assessment of structural similarity between the docking results and established templates.

CONCLUSION

The analysis of molecular interactions is fundamental to understanding the formation and stability of ligand-protein complexes. Large-scale analysis of many complexes becomes repetitive. An alternative is the analysis of vector representations of the interactions. NCS offers a solution to generate and analyze interaction vectors. Its flexibility allows for analysis with different focus.

Acknowledgement

This project is supported by the Brazilian agency CNPQ.

Conflicts Of Interest

There are no conflicts of interest

REFERENCES:

[1] Bender, A., Glen, R.C. (2004) Molecular similarity: a key technique in molecular informatics, Org. Biomol. Chem., 2, 32043218.
[2] Deng, Z., Chuaqui, C., Singh, J. (2004) Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions, J. Med. Chem., 47, 337344.
[3] Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J. Chem. Inf. Model., 61(8), 3891–3898.
[4] Jain, A.N. (2003) Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine, J. Med. Chem, 46, 499-511.
[5] Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. (2007) ClustalW and ClustalX version 2, Bioinformatics, 23, 2947-2948.

[6] McDonald, I.K., Thornton, J.M (1994) Satisfying hydrogen bonding potential in proteins, J. Mol. Biol., 238, 777793.
[7] Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, J. Comput. Chem., 30, 1-1.
[8] Desaphy, J., Raimbaud, E., Ducrot, P., & Rognan, D. (2013). Encoding protein-ligand interaction patterns in fingerprints and graphs. J. Chem. Inf. Model., 53(3), 623–637. Rognan, D. (2001) Article title, Journal Name, 99, 33-54.
[9] Sobolev,V., Sorokine, A., Prilusky, J., Abola, E., Edelman, M. (1999) Automated analysis of interatomic contacts in proteins, Bioinformatics, 15, 327-332.
[10] Xue, L., Golden, J.W., Stahura, F.L., Bajorath, J. (2003) Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys, . J. Chem. Inf. Model., 43, 1218-1225.