Strategies for the Effective Identification of Remotely Related Sequences in Multiple PSSM Search Approach

V.S. Gowri, K.G. Tina, O. Krishnadev, and N. Srinivasan*
Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

ABSTRACT **Searches using position specific scoring matrices (PSSMs) have been commonly used in remote homology detection procedures such as PSI-BLAST and RPS-BLAST. A PSSM is generated typically using one of the sequences of a family as the reference sequence. In the case of PSI-BLAST searches the reference sequence is same as the query. Recently we have shown that searches against the database of multiple family-profiles, with each one of the members of the family used as a reference sequence, are more effective than searches against the classical database of single family-profiles. Despite relatively a better overall performance when compared with common sequence-profile matching procedures, searches against the multiple family-profiles database result in a few false positives and false negatives. Here we show that profile length and divergence of sequences used in the construction of a PSSM have major influence on the performance of multiple profile based search approach. We also identify that a simple parameter defined by the number of PSSMs corresponding to a family that is hit, for a query, divided by the total number of PSSMs in the family can distinguish effectively the true positives from the false positives in the multiple profiles search approach. Proteins 2007;67:789–794. © 2007 Wiley-Liss, Inc.**

Key words: position specific scoring matrix; protein profiles; protein families; remote homology detection; sequence analysis

INTRODUCTION

Genome sequencing projects have resulted in the availability of amino acid sequences of several thousands of proteins. The number of entries of amino acid sequences of proteins in SWISSPROT is over 218,000.[1] Although the databank of known structures of proteins contains over 36,000 entries, the number of distinct proteins of known structure is much less. This emphasizes the known trend of increase in sequence space is much higher than the increase in structural space. One of the approaches to bridge the gap between the sequence space and structural space is to identify the relationships between these genome-derived amino acid sequences and the proteins of known structure.

It is also very common that many proteins predicted from genomic data have no known function and prediction of function by homology search is not always successful. One of the common underlying problems in prediction of function and structure by homology search is the inability to identify correctly the related proteins when the sequence similarity is poor. It is very common that due to high divergence of sequences during evolution, they share poor sequence identity ($< \sim 30\%$). Under such a circumstance simple pairwise comparison procedures often fail to relate the distantly related sequences with their distantly related homologues.

Profile based procedures such as PSI-BLAST,[2] HMMER,[3–6] and RPS-BLAST[7,8] have been shown to identify the relationships between protein families that are distantly related. However, in PSI-BLAST and RPS-BLAST a single PSSM is used corresponding to a given multiple sequence alignment. PSSM integrates the extent of occurrence of each of the 20 residue types and the probability of replacement of a residue in the reference sequence by each one of the 20 residue types. In PSI-BLAST searches, the query sequence is generally used as the reference sequence for PSSM generation. In RPS-BLAST searches, the reference sequence is chosen arbitrarily. Hence, the PSSMs generated using different reference sequences will be different. It has been shown convincingly that the sensitivity of the PSSM depends on the choice of the reference sequence.[9]

Recently, we have shown that the use of multiple PSSMs corresponding to a multiple sequence alignment improves substantially the effectiveness of the remote homology detection.[9] We have shown that given a multiple sequence alignment of n sequences, one can generate n PSSMs, each corresponding to use of one of the sequences in the alignment as a reference sequence.[9] A comparison of searches against the databases of single PSSMs, HMMs and multiple PSSMs shows that search against the database of multiple PSSMs has better sensitivity, specificity,

Grant sponsors: Council of Scientific and Industrial Research, New Delhi; Wellcome Trust, UK; Department of Biotechnology, India

*Correspondence to: N. Srinivasan, Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India. E-mail: ns@mbu.iisc.ernet.in

Received 2 July 2006; Revised 14 October 2006; Accepted 27 November 2006

Published online 22 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21356

and less error rate when compared with searches against databases of HMMs or single PSSMs.[9] Searches against multiple PSSM databases using RPS-BLAST are also computationally more economical when compared with searches against HMMs. A database of multiple PSSMs[10] has been generated for the sequence and structural protein domain families. The RPS-BLAST interface enables the user to query the PSSM databases.[10] This database is available in the public domain at http://crick.mbu.iisc.ernet.in/∼mulpssm.

Our earlier analysis on the comparison of performance of single PSSMs with multiple PSSMs showed that the number of false positives resulting from multiple PSSM approach were very few and most of them are due to local structural similarity or due to short lengths of the PSSMs.[9] In this article, we discuss the practical ways of distinguishing true and false positives for the searches against multiple PSSM databases. We have also analyzed the effect of sequence identity based clustering and the effect of profile length on the performance of multiple PSSM search approach.

MATERIALS AND METHODS

Dataset

For the purposes of assessment and comparisons, we chose to use the structural members from PALI database,[11] which is derived from SCOP database,[12] and hence the relationship between protein domains and families of protein domains in the database is already known. Searches have been made against the PSSM databases using the structural members from PALI database as queries. The PSSMs of the protein families have been generated using a set of integrated sequence-structure alignments obtained using an approach described previously.[13] Briefly, the approach involves generation of a database of PSSMs for the PFAM[14,15] families. Then the corresponding PALI and PFAM families are unified. For this purpose, every sequence in the PALI families are queried against the PFAM family profiles using RPS-BLAST (*E*-value cut-off: 0.00001). For protein families that cannot be unambiguously related to the PFAM families, searches are made with queries from such PALI families against NRDB using PSI-BLAST (*E*-value cut-off: 0.0001). These searches against PFAM PSSMs or NRDB results in sets of homologous sequences for a given PALI family from PFAM or NRDB. For each of these sets of homologues, an integrated sequence-structure alignment and the corresponding PSSM are generated using PSI-BLAST.

The proteins used in the analysis were obtained from the PALI database (Release 2.1), which contains 518 families with at least 3 members in each family. These protein families were further filtered using a sequence identity cut-off of 60% between the members in a family in order to identify very closely related proteins and retain only one representative for the very closely related (greater than 60% sequence identity) homologues. The final dataset used contains 1325 sequences from 286 multimember families. The integrated-sequence structure alignments for these 286 multimember families used in this analysis are available publicly at http://crick. mbu.iisc.ernet.in/∼mulpssm/iss_db.

Generation of PSSMs

Both the single and multiple PSSMs are generated for all the 286 families. There are two kinds of single PSSM databases generated. Single PSSM databases are generated using both the shortest sequence and longest sequence in every family as reference sequences. For the purposes of studying the effect of profile length on the performance of multiple PSSM approach, nine multiple PSSM databases are generated using reference sequences chosen at different length cut-offs ranging from 20% to 90% of the length of the longest sequence in each family. RPS-BLAST searches with all 1325 structural members are performed against both the single and multiple PSSM databases individually. Sensitivity, specificity, and error rates (%) are calculated as described below:

(i) Sensitivity = TP/(TP + FN)
(ii) Specificity = TP/(TP + FP)
(iii) Error rate = (FP + FN)/TP

Where TP, FP, and FN correspond to number of true positives, false positives, and false negatives, respectively.

Clustering of Sequences for PSSM Generation

To study the effect of extent of sequence similarity on the sensitivity of multiple PSSMs for a family, we generated PSSMs for a family at different sequence identity cut-offs. 100 families were chosen at random from the PALI database and the sequences of each of the families were clustered at 20–90% identity cut-offs at every 10%. Clustering of the sequences within a family was done using BLASTCLUST program.[16] From each cluster for a given family, the longest sequence was chosen to represent the cluster and PSSMs were generated as described already.

RESULTS AND DISCUSSIONS

Identification of False Positives from the Hits Obtained from Searches against Multiple PSSM Databases

In this section, we suggest a simple parameter to aid identification of false positives from a set of PSSM hits in the multiple profile-based sequence search. Although the initial set of hits are scrutinized through filters such as E-value and query coverage, small number of false positives occur.[9] The reasons for the false positive connections are often local structural similarity between the query and PSSM and shorter lengths of the profiles.[9] To distinguish the true positives from the false positives in multiple PSSM approach, we calculated the Percentage PSSM Factor (PP_{Factor}), which is defined below, for every search

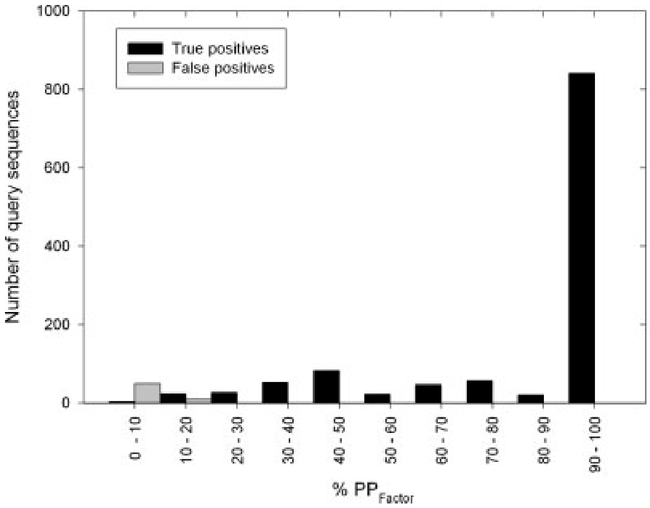

Fig. 1. % PP_{Factor} Vs Number of query sequences for searches against multiple family profiles.

in the database of multiple profiles:

$$PP_{Factor} = (N_i/N) \times 100$$

Here, N_i is the number of profiles hit for a query in a family ’i’; N is the total number of profiles in the family ’*i*’ present in the PSSM database. We explored if a large PP_{Factor} signifies the reliability of association of the query sequence with the family ‘*i*’. An analysis has been made on the results of multiple PSSMs search for queries of known structure (from PALI database) searched against a database of multiple PSSMs of PALI families. The distribution of number of query sequences with various PP_{Factor} values is given in Figure 1. The PP_{Factor} for a true positive falls in the range of 0–100% with 98% of the true positives showing a PP_{Factor} above 20%. The highest PP_{Factor} for a false positive is 20%. This shows that, if a query sequence matches with large number of profiles for a family (greater than 20%), then, the relationship between the query sequence and the hit family is highly reliable. When the query sequence matches with few profiles (less than 20%) for a family with significant e-values, then such connections have a good chance of being false positives.

Comparison of Performances of Searches Against Profiles with Varying Lengths

The performance of searches against single profiles using the longest sequence as the reference sequence has been already compared.[9] However, it can be argued that the use of shortest sequence in a family as the reference sequence has certain advantages as it results in the shortest PSSM possible for the family. It is important to consider the performance of profile matching using PSSMs of various profile lengths as the performance is also dependent on the ability of the profile matching program to generate optimal or sub-optimal alignments when the lengths of the profile and the query are substantially different. Hence, we generated RPS-BLAST searchable database of single profiles using shortest sequence in every family, apart from using longest sequence, as the reference sequence. During our searches against multiple family profiles, a few known relationships are missed (false negatives) by our search procedure. Hence, we performed an analysis in which searches are made against databases of multiple PSSMs generated using reference sequences of varying lengths. The reference sequence lengths in these databases vary from 20 to 90% of the length of the longest sequence for every family. Parameters such as sensitivity, specificity, and error rates (%) are calculated for searches against single PSSMs and multiple PSSM databases and are discussed as follows:

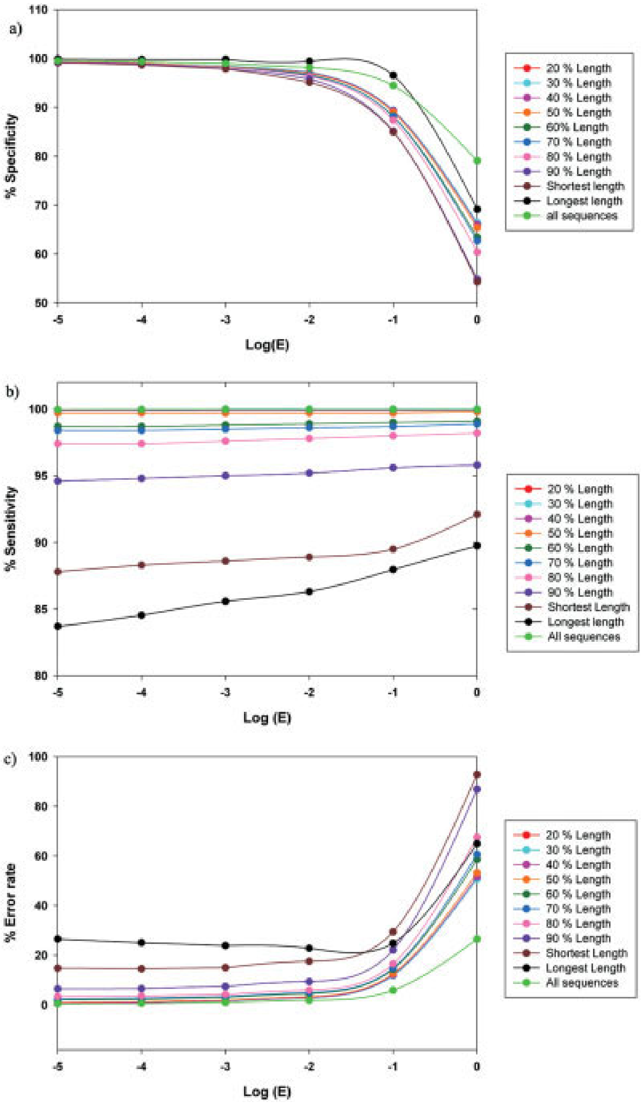

Fig. 2. Plots of performance indicators against Log (E) for searches against the databases of single and multiple PSSMs. (a) % Specificity, (b) % Sensitivity, (c) % Error rate.

Specificity

A plot of % specificity as a function of log (E) value (Fig. 2a) shows that at stringent E-value cut-offs of 10^{-5} to

10^{-3}, the specificities are nearly comparable and independent of the length of the reference sequences used. As the E-value cut-off relaxes from 10^{-2} to 1, the shortest sequence has the poorest specificity and the longest has a better specificity. However, the multiple profiles containing a mixture of profiles with different lengths perform better than any of the other profile databases.

Sensitivity

A plot of % sensitivity as a function of log (E) value (Fig. 2b) for the searches against a set of family PSSM databases with varying lengths of reference sequences shows that as the lengths of the reference sequences decreases from about 90 to 20% of the length of the longest sequence, the sensitivity increases as high as 5%. In contrast to the specificity, the searches against PSSM database using shortest sequence have good sensitivity when compared with PSSMs generated using longest sequence. Use of all the sequences also results in the best performance in terms of sensitivity.

Error rate

The Error rate (%) is plotted as a function of Log (E) value for all the searches made against the database of PSSMs generated using reference sequences at various length cut-offs (Fig. 2c). This plot shows that the shortest profiles have the lower error rates at stringent E-values; however, at relaxed E-values of 10^{-1} to 1, the single PSSMs generated using longest length sequences perform better and have low error rate values. In the case of searches against database of shortest profiles, the error rate is low at stringent E-value cut-off (0.00001) as we have employed a profile coverage cut-off of 70% in order to filter the false hits. Hence, at stringent E-value cut-offs, the probability of a database entry related to the query passing the coverage cut-off is higher with a shorter profile. However, searches against multiple PSSMs with all the structural members as reference sequences have the lowest error rate at all the E-values and the decrease in error rate is 25–60%. In the case of sets of profiles with lengths shorter than the longest profile data set, the error rate is lower at all the E-value cut-offs. This is due to the fact that searches against profiles with shorter lengths reduce the number of false negatives, thus reducing the error rate.

Comparison of the Extents of True Positives, False Positives, and False Negatives as a Function of Lengths of the Reference Sequences

Though the analyses discussed earlier show clearly the effect of lengths of the reference sequences on the nature of the PSSM databases generated, we further analyzed variation in the number of true positives (expressed as %), false positives, and false negatives as function of lengths of the reference sequences at the stringent E-value cut-off of 10^{-5}. A plot of percentage of true positives, false positives, and false negatives as a function of the length of profiles (Fig. 3) indicates that as the length of the profiles decreases, the rate of false positives increases and remains constant after reaching a threshold. However, the percentage of false negative decreases gradually as the length cut-off decreases. There is a sudden increase in the percentage of true positives in searches from single PSSMs to multiple PSSMs. In general, there is a gradual increase in the percentage of true positives as the length of PSSMs decreases.

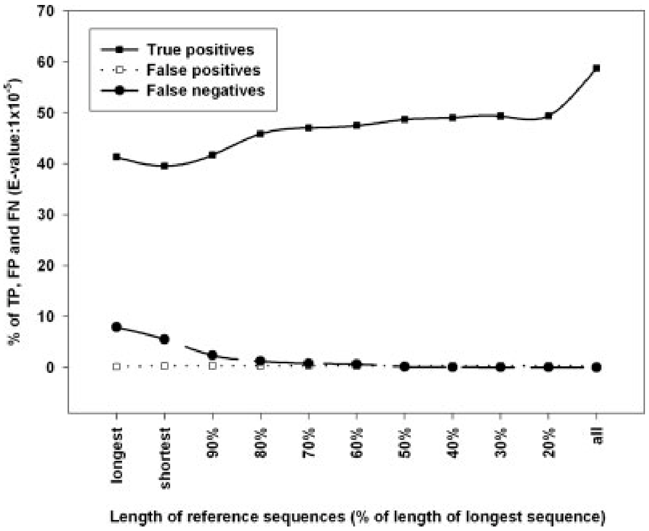

Fig. 3. Plot of % of true positives (TP), false positives (FP), and false negatives (FN) as a function of length of the profiles in single and multiple profiles database at E-value of 10^{-5}.

Performance Comparison of Multiple Profiles Approach based on Clustering of Sequences

The elements of a column in a PSSM generated by PSI-BLAST (which corresponds to a position in alignment) are biased towards the amino acid present in the query used, at that position. Thus, in the case of two very similar reference sequences the PSSMs would be very similar. Therefore, generating PSSMs for all the members of a given family is not a very efficient strategy as the information in such a database will have some redundancy. If the number of PSSMs becomes very large searching such a database would be computationally expensive. To study the effect of homology on the sensitivity of multiple PSSMs for a family, we generated PSSMs for a family at different clustering cutoffs. Hundred families were chosen at random from the SCOP/PALI database and the sequences of each of the families were clustered, as given in detail in Methods section, at different identity cutoffs to remove redundancy. From each cluster for a given family, the longest sequence was chosen to represent the cluster and PSSMs were generated as described in the Methods section. As expected, the number of PSSMs generated was drastically reduced at lower clustering cutoffs. The number of true positives and false positives obtained at each clustering cutoff is given in Table I, which also gives the number of PSSMs generated at each clustering cutoff and the time required per sequence to search in the database.

TABLE I. The Number of True Positives and False Positives along with the Sensitivity and Specificity of Remote Homology Detection with PSSMs Generated at Different Clustering Cutoffs

Identity cutoff for PSSM generation	Number of true positives	Number of false positives	Specificity (%)	Sensitivity (%)	Total number of PSSMs generated	Time (s) taken for querying 1 sequence against the database[a]
20	37	22	62.71	56.92	1394	0.2
30	37	25	59.68	56.92	1420	0.2
40	44	32	57.89	67.69	1990	0.35
50	48	33	59.26	73.85	4321	0.8
60	54	35	60.67	83.08	8654	1.5
70	59	35	62.77	90.77	14339	2.7
80	61	36	62.89	93.85	20980	4.75
90	65	37	63.73	100	23785	5.55

The number of PSSMs generated at each clustering cutoff is also given in the table. The sensitivity value given is calculated as the percentage of number of hits detected at 90% clustering cutoff in order to show the reduction in sensitivity after lowering the clustering cutoff.

[a] The values correspond to average time taken for 10 runs with 20 different sequences.

From the data given in Table I it can be seen that as the sequences considered for generation of PSSMs become more divergent, the sensitivity decreases while the specificity does not show any correlation with the clustering cutoff. The increase in the sensitivity is approximately linear with increase in clustering cutoff (R^2 value: 0.98). In contrast to the linear relationship between the sensitivity and clustering cutoff, there is an exponential increase in the number of profiles and consequently the time required to search in the database with increase in the clustering cutoff (R^2 value 0.86 with time proportional to the clustering cutoff raised to the power 1.8). Thus, the computational expense for searching a large database is not compensated by a proportional increase in the sensitivity. It must be pointed out that the time required for searching the database is dependent on the computational resources available and the values given in Table I are for a machine with Intel P4 3 GHz processor (512 MB Cache) with 1GB RAM.

If the availability of computational resources is not limited or the time required for searching large databases is not a constraint, then generating profiles at a clustering cutoff of 100% is the best choice as such a database will have the maximum sensitivity.

CONCLUSIONS

Our analysis on the identification of protein families related to a query, from the hits obtained from the multiple PSSM approach, shows that when a query sequence matches with almost all of the multiple profiles of a given family, then the relationship between the query sequence and the hit family is highly reliable and if the query sequence matches with only few profiles of a family then, such a relationship is probably a false positive. This is further quantified by calculating the PP_{Factor} for every family hit by the query sequence. Further, searches against the database of shortest family profiles have been shown to result in substantial proportion of false positive hits and those against longest sequence profiles results in substantial proportion of false negatives. However, searches against the database of multiple family profiles generated using every member in the family as reference sequence shows best performance in terms of sensitivity, specificity, and error rate when compared with single and multiple family profiles generated at various length cut-offs. In addition to the lengths of the profiles, PSSM generation using nonredundant sequences as reference sequences have been proved to be sensitive and minimize the computational search time. It is suggested that, as far as possible, the clustering cutoff of 100% is used as the sensitivity of the profile database is highest at this cutoff and there is no redundancy in the PSSMs generated.

REFERENCES

1. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res 1998;26:38–42.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
3. Baldi P, Chauvin Y, Hunkapiller T, McClure MA. Hidden Markov models of biological primary sequence information. Proc Natl Acad Sci 1994;91:1059–1063.
4. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modelling. J Mol Biol 1994;235:1501–1531.
5. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
6. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
7. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 1999;12:1000–1011.
8. Marchler-Bauer, A, Panchenko, AR, Shoemaker, BA, Thiessen, PA, Geer, LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 2002;30:281–283.
9. Anand, B, Gowri, VS Srinivasan N. Use of multiple profiles corresponding to a sequence alignment enables effective detection of remote homologues. Bioinformatics 2005;21:2821–2826.

10. Gowri VS, Krishnadev O, Swamy CS, Srinivasan N. MulPSSM: a database of multiple position-specific scoring matrices of protein domain families. Nucleic Acids Res 2006;34:D243–D246.
11. Balaji S, Sujatha S, Kumar SS, Srinivasan N. PALI—a database of Phylogeny and ALIgnment of homologous protein structures. Nucleic Acids Res 2001;29:61–65.
12. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
13. Gowri VS, Pandit SB, Karthik PS, Srinivasan N, Balaji S. Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. Nucleic Acids Res 2003;31:486–488.
14. Sonnhammer, EL, Eddy, SR, Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 1997;28:405–420.
15. Bateman, A, Coin, L, Durbin, R, Finn, RD, Hollich, V, Griffiths-Jones, S, Khanna, A, Marshall, M, Moxon, S, Sonnhammer, ELL et al. The Pfam Protein Families Database. Nucleic Acids Res 2004;32: D138–D141.
16. Wheeler, DL, Barrett, T, Benson, DA, Bryant, SH, Canese, K, Church, DM, DiCuccio, M, Edgar, R, Federhen, S, Helmberg, W. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2005;33:D39–D45.