ISSN 1330-9862
doi: 10.17113/ftb.55.02.17.4749

scientific note

MEGGASENSE – The Metagenome/Genome Annotated Sequence Natural Language Search Engine: A Platform for the Construction of Sequence Data Warehouses

*Ranko Gacesa[1,2], Jurica Zucko[1,3], Solveig K. Petursdottir[4], Elisabet Eik Gudmundsdottir[4], Olafur H. Fridjonsson[4], Janko Diminic[1,3], Paul F. Long[2,5], John Cullum[6], Daslav Hranueli[1,3], Gudmundur O. Hreggvidsson[4,7] and Antonio Starcevic[1,3]**

[1]SemGen Ltd., Lanište 5/D, HR-10 000 Zagreb, Croatia

[2]Institute of Pharmaceutical Science, King's College London, Franklin-Wilkins Building, Stamford Street, London SE1 9NH, UK

[3]Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, HR-10 000 Zagreb, Croatia

[4]Matis Ltd., Vínlandsleið 12, IS-113 Reykjavík, Iceland

[5]Faculty of Life Sciences and Medicine, King's College London, Franklin-Wilkins Building, Stamford Street, London SE1 9NH, UK

[6]Department of Genetics, University of Kaiserslautern, Postfach 3049, DE-67653 Kaiserslautern, Germany

[7]Faculty of Life and Environmental Sciences, University of Iceland, Sturlugötu 7, IS-101 Reykjavík, Iceland

Received: April 22, 2016
Accepted: January 17, 2017

Summary

The MEGGASENSE platform constructs relational databases of DNA or protein sequences. The default functional analysis uses 14 106 hidden Markov model (HMM) profiles based on sequences in the KEGG database. The Solr search engine allows sophisticated queries and a BLAST search function is also incorporated. These standard capabilities were used to generate the SCATT database from the predicted proteome of *Streptomyces cattleya*. The implementation of a specialised metagenome database (AMYLOMICS) for bioprospecting of carbohydrate-modifying enzymes is described. In addition to standard assembly of reads, a novel 'functional' assembly was developed, in which screening of reads with the HMM profiles occurs before the assembly. The AMYLOMICS database incorporates additional HMM profiles for carbohydrate-modifying enzymes and it is illustrated how the combination of HMM and BLAST analyses helps identify interesting genes. A variety of different proteome and metagenome databases have been generated by MEGGASENSE.

Key words: bioprospecting, carbohydrate-modifying enzymes, DNA assembly

*Corresponding author: Phone: +385 1 4605 147; Fax: +385 1 4836 083; E-mail: astar@pbf.hr

ORCiD IDs: 0000-0003-2119-0539 (Gacesa), 0000-0001-7782-6503 (Zucko), 0000-0003-4033-3654 (Petursdottir), 0000-0002-3404-850X (Gudmundsdottir), 0000-0002-8725-602X (Fridjonsson), 0000-0001-5104-5813 (Diminic), 0000-0001-6410-5803 (Long), 0000-0002-3850-8526 (Cullum), 0000-0001-8336-4384 (Hranueli), 0000-0002-4958-1673 (Hreggvidsson), 0000-0003-2386-2124 (Starcevic)

Introduction

Falling costs of next generation sequencing have made *de novo* genome and metagenome sequencing widely available. After assembly of the sequencing reads, a genome is represented by a large number of contigs with the presence of many gaps; the gaps arise from DNA regions which are difficult to sequence (*e.g.* recalcitrant to PCR) and from assembly problems (*e.g.* the presence of repeated sequences). The same problems occur in a metagenome, but are exacerbated by the presence of different species often in greatly different proportions (*i.e.* some rarer species may only have low level coverage). Bioinformatics offers many tools to analyse the sequences, and the identification of protein-coding regions and assignment of function are the major aim in most projects. For metagenomes, a phylogenetic analysis of the present species is usually the second important aim.

There are effective statistically based methods to identify probable protein-coding regions (*1*). There are also many tools to try to assign function to such proteins. The BLAST algorithm (*2*) will detect similar sequences in databases. A general BLAST database such as GenBank (*3*) consists mainly of uncurated entries, which will often contain misleading data for functional assignment. The SEED database (*4*) contains collections of protein sequences grouped by function and has been used for BLAST searches to find hits corresponding to *in silico* translation of the metagenomic sequences. Once hits are found, they can be used to assign function using comparison to FIGfams protein families (*5*). In order to present functional information about the whole genome or metagenome effectively, it is necessary to have a suitable data structure. The KEGG database (*6*) has developed the BRITE classification, which is a hierarchical scheme to incorporate different functions (often enzyme activities). There is a considerable quantity of curated information about each class and links to pathway diagrams. The KEGG database has a collection of KEGG orthologues associated with each functional class. BLAST searches against the KEGG orthologues are a useful way of assigning function to new sequences.

A general weakness of BLAST analyses is that all regions of the compared proteins are given equal weighting. In contrast, a hidden Markov model (HMM) profile constructed from a family of proteins assigns higher weighting to important conserved residues (*7*). The Pfam database (*8*) contains HMM profiles of protein families and can be used to assign function to sequences. There are also many specialised analyses, which can be used for particular projects, *e.g.* carbohydrate-metabolising enzymes (CAZy database (*9*)). Phylogenetic analysis of metagenome sequences is fairly effective. The Phymm system (*10*) uses coding sequences for this purpose and has been incorporated in the Glimmer-MG system (*1*) to identify protein-coding regions in metagenomic sequences.

The biggest problem of sequence annotation is to present the results in a form that is useful for biological scientists. Analysis of a genome or a metagenome generates an overwhelming amount of information and it is difficult to extract the relevant data. An elegant solution is to enter the data into a suitable database with appropriate search, analysis and extraction functions. The MG-RAST database (*11*) is designed for metagenome sequences and new data are analysed using a standard pipeline. Experience with different genome and metagenome analyses showed that the required analyses and tools differed between different projects and it was often necessary to carry out specialised analyses for particular projects. This makes use of a standard database difficult. However, most projects required common components. It was, therefore, decided to develop the MEGGASENSE platform, which would allow the generation of different databases for different projects. The databases can be constructed from DNA sequencing reads, assembled DNA sequences or protein sequences. A core functional analysis common to all databases uses HMM profiles constructed from KEGG orthologues (KO). Any other required analyses can be incorporated (*e.g.* BLAST-based). Metagenome samples are also subjected to a phylogenetic analysis. The whole of the annotations (functional and phylogenetic) can be searched using a powerful search engine (the default is Solr) and it is easy to extract individual sequences and sets of sequences from the databases. The utility of MEGGASENSE is illustrated by some example databases.

Materials and Methods

The databases are implemented in ZODB (*12*). The search engine is the enterprise search platform Lucene/Solr (*13*) served by a Tomcat web server (*14*) to index and serve the annotated sequence libraries. The web logic was implemented by an HTML (*15*) and a JavaServer Pages (JSP) (*16*) combination. The databases also include BLAST+ v. 2.2.25 (*17*) and HMMER v. 3.0 (*18*) search functions. Metagenomic DNA databases also incorporated the Krona viewer (*19*) for phylogenetic data.

The KEGG database v. 58 (*6*) was downloaded and the annotation associated with each KEGG orthologue (KO) was retrieved with custom database loading programs. An HMM profile was built from the sequences of each of the 14 106 KOs using HMMER v. 3.0 (*18*) after multiple alignment of the sequences using ClustalW (*20*).

Optional additional analyses can be added. For the AMYLOMICS database, more detailed analysis of carbohydrate-modifying enzymes was carried out using HMM profiles from dbCAN (http://csbl.bmb.uga.edu/dbCAN/ (*21*)). These profiles correspond to different enzyme families classified in the CAZy database (http://www.cazy.org (*9*)). Another optional analysis step is identification of reads belonging to ncRNAs, for which we use Infernal (http://eddylab.org/infernal/ (*22*)) coupled with a local implementation of the Rfam database (http://rfam.xfam.org/ (*23*)).

DNA assembly used the Newbler package, which is designed specifically for the 454 GS series of pyrosequencing platforms sold by Life Sciences, a Roche Diagnostics company (*24*). In the conservative annotation approach, the DNA reads were assembled and the resulting contigs were translated in all six reading frames and screened with the KEGG-based HMM profiles to generate the default functional annotation. The functional annota-

tion approach first screened the reads using a customised version of the Glimmer-MG v. 0.2 pipeline (*1*) to identify potential protein-coding regions. This pipeline includes Scimm, PhymmBL and related tools (http://www.cbcb.umd.edu/software/scimm/ (*25*)), which also produce the information about the phylogeny of the potential coding regions in the metagenomic samples. The protein-coding regions were screened with the KEGG-based HMM profiles and the sequences corresponding to each KO were grouped together and assembled using the Newbler assembler. The results of the conservative and functional annotations were merged with duplicates recognised using BLAST (*17*). Additional specialised analyses (*e.g.* the CAZy analysis for AMYLOMICS) result in additional annotation detail, which was associated with each entry.

The MEGGASENSE platform was initially developed as part of a bioprospecting project for carbohydrate-modifying enzymes using metagenomes from hot springs in Iceland. The metagenomic libraries used to construct the AMYLOMICS database (http://mrcina.bioinfo.pbf.hr/Amylomics/) were derived from samples collected from three locations at geothermal sites in Iceland and subjected to carbohydrate enrichment using α- and β-glucan substrates. The metagenome MET4 was collected at Hærulangur (Vonarskarð geothermal region, Iceland) from the effluent (dark layer beneath white mat, 72 °C, pH=6) and enriched on cotton whisk as carbon source.

The HMM profiles based on KEGG and the custom programs for MEGGASENSE are available on github (https://github.com/astarsky2016/meggasense).

Results

Simple proteome databases

The MEGGASENSE platform is designed to generate sequence databases from different types of input data: raw DNA sequencing reads, finished DNA sequences or proteome sequences. Database construction involves several steps (Fig. 1). In the first step, the sequences are collected and annotated. The default functional analysis for databases generated by MEGGASENSE is derived from the KEGG database. We constructed HMM profiles from 14 106 KOs. Each database consists of a data warehouse containing various data sources: relational databases, object databases (implemented in ZODB) and files. The core functions of MEGGASENSE can be illustrated by the construction of simple proteome databases.

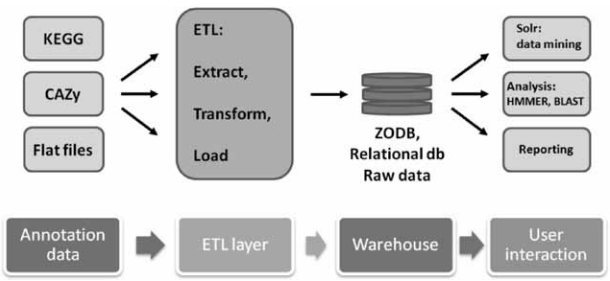

Fig. 1. Database construction using the MEGGASENSE platform. Different analysis tools are used to analyse the input sequences contained in flat files. The data and the results of analyses are used to construct the data warehouse. The graphical user interface allows searching using the Solr search engine (*13*) as well as other analyses and reporting

The SCATT database (http://bioserv7.bioinfo.pbf.hr/scattDB/registration/login.jsp) was constructed from the 6949 protein sequences in the predicted proteome of *Streptomyces cattleya* (*26*). The 14 106 HMM profiles were used to scan the proteome to associate the proteins with the KOs. The protein sequences and the annotations associated with each KO were loaded into the new database using a custom database loading program (Fig. 1). The graphical user interface includes search functions using the Solr search engine and the ability to generate tables using the KEGG BRITE classification. This makes it easy to investigate which metabolic pathways are present in the strain and to identify any missing steps in pathways. The database also implements BLAST and HMMER searches (Fig. 1). A similar process was used to construct the ZoophyteBase database (http://bioserv7.bioinfo.pbf.hr/Zoophyte/registration/login.jsp) from the proteome of the coral *Acropora digitifera* (*27*). The search functions and the hierarchical KEGG BRITE classification allow the recognition of particular functions as well as collections of functions belonging to particular pathways. This enabled the generation of some novel hypotheses about functions which may be involved in coral symbiosis.

Metagenome databases with additional specialised functions

The starting point for a metagenome database is usually raw DNA sequencing reads. This means that in the data and annotation phase of construction (Fig. 1) it is necessary to carry out assembly and gene identification. Phylogenetic analyses are also carried out at this stage using the Phymm analysis in the Glimmer-MG pipeline (*1*) and from the detected 16S rRNA sequences. The graphical user interfaces of metagenome databases generated by MEGGASENSE incorporate the Krona tool for viewing the phylogenetic results. For data mining projects, it is often useful to add further specialised analyses.

The AMYLOMICS database used the 454 sequencing technology and the reads were assembled using the Newbler assembler (*24*). After *in silico* translation of the DNA sequences, gene function was assigned using the KEGG KO HMM profiles as described earlier. Assembly of metagenome sequences is considerably more difficult than of genome sequences, because some of the species are present in low numbers so that there are many regions with low read coverage. A novel 'functional' assembly strategy was also implemented to improve the assembly of such sequences. This used a pre-screening of the reads with the HMM profiles followed by an attempt to assemble genes. The reads were extracted from the files produced by the sequencer and trimmed to remove low-quality sequences. The Glimmer-MG pipeline (*1*) was used to detect protein-coding regions. After *in silico* translation, the reads were screened with the HMM profiles derived from the KEGG database and reads corresponding to each KO were collected together. It was necessary to implement an algorithm to resolve cases where more than one profile gave a hit with a single read. In some cas-

es, this was due to artefacts with low scores, whereas in other cases, it was because the read overlapped two genes. The reads corresponding to each KO were then assembled into contigs using the Newbler assembler. The results of the traditional and functional analyses were merged: BLAST was used to compare the two sets of results and eliminate duplicates. Table 1 shows a comparison of the traditional and functional approaches to assign genes to a KEGG BRITE functional category for a metagenome sample in the AMYLOMICS database. The 'functional' assembly strategy added about 35 % more predicted genes represented by contigs with two or more reads. Reads that could not be assembled were retained in the database as singleton reads. The MG-RAST pipeline is a popular tool for analysing metagenome data (*11*). It can also be used to assign genes to KEGG BRITE categories. The sequences of the metagenome were submitted to the MG-RAST server. Table 1 shows that MG-RAST assigned far fewer genes to KEGG BRITE categories than the AMYLOMICS database.

Table 1. KEGG BRITE categories of hits identified in the metagenomic sample MET4 in the AMYLOMICS database (http://mrcina.bioinfo.pbf.hr/Amylomics/)

KEGG functional categories	Number of identified genes		
	Traditional	Functional	MG-RAST
Cellular processes	**299**	**454**	**139**
Transport and catabolism	31	75	14
Cell motility	221	301	94
Cell growth and death	46	70	31
Cell communication	1	8	0
Metabolism	**5626**	**7329**	**1250**
Carbohydrate metabolism	1269	1553	341
Energy metabolism	719	1013	99
Lipid metabolism	226	340	44
Nucleotide metabolism	596	677	122
Amino acid metabolism	931	1052	382
Metabolism of other amino acids	92	96	11
Glycan biosynthesis and metabolism	331	528	63
Metabolism of cofactors and vitamins	628	683	141
Metabolism of terpenoids and polyketides	119	173	23
Biosynthesis of other secondary metabolites	15	25	23
Xenobiotics biodegradation and metabolism	62	111	1
Environmental information processing	**2170**	**3481**	**360**
Membrane transport	1656	2449	263
Signal transduction	428	698	97
Signalling molecules and interaction	86	334	0
Genetic information processing	**2753**	**3600**	**425**
Transcription	741	1053	35
Translation	724	857	192
Folding, sorting and degradation	384	646	71
Replication and repair	904	1044	127

The analyses described above are incorporated as a default in metagenome databases generated by MEGGASENSE. In addition to the default annotation based on the KEGG database, the AMYLOMICS database used specialised HMM profiles for better identification of carbohydrate-modifying enzymes. These were obtained from the dbCAN database (*21*) and are based on sequences in the CAZy database (http://www.cazy.org (*9*)), which includes sequences of carbohydrate-modifying enzymes arranged in different families. The bioprospecting approach aimed at identifying gene sequences encoding enzymes with novel properties. The AMYLOMICS database contains 279 and 568 putative carbohydrate-modifying enzymes for samples enriched on α- and β-glucans, respectively, but it is likely that some hits with low scores are falsely identified. This was examined by using the BLAST function of the AMYLOMICS database, followed by sorting the results according to the degree of sequence identity with the best BLAST hit for each sequence (Fig. 2). Most of the BLAST hits were annotated as carbohydrate-modifying enzymes. However, for sequences with a very low degree of sequence identity (25–30 %), the BLAST hits were usually not annotated as carbohydrate-degrading enzymes (Fig. 2a). Examination of the alignment of these sequences with the HMM profiles to see which residues were responsible for the functional assignment decided whether the sequence would be rejected as a probable artefact or retained as a potential novel enzyme. This reduced the number of potential enzymes to 250 and 519, respectively (Table 2). Some sequences were nearly identical to the known sequences and considered unlikely to yield novel enzymes (Fig. 2c). The majority of the sequences (Fig. 2b) showed a lower degree of sequence identity and are being analysed in more detail using biochemical knowledge about the enzyme families.

Table 2. Categories of enzymes using carbohydrate substrates identified in metagenomic samples from hot springs in Iceland

Enzyme function	β-glucan enrichment	α-glucan enrichment
1,4-α-glucan branching enzyme	55	12
α-amylase	72	48
α-glucosidase	25	12
α-mannosidase	83	23
β-amylase	2	7
cyclodextrinase	48	26
cyclomaltodextrin glucanotransferase	24	21
glucoamylase	21	20
glycogen-debranching enzyme	15	6
maltogenic amylase	14	18
neopullulanase	72	27
pullulanase	88	30
Total	519	250

a)

Genes of Interest (filtered): (424) Original Table Download Sequences Display Sequences Detailed Display Reset Select All

seq_id	assigned annotation	most similar to annotated	percent identity to most similar database hit	status	remarks
3359647	Maltogenic amylase	septation ring formation regulator EzrA [Mycoplasma sp. CAG:776]	25.47	singleton	functional approach
3200138	Alpha-mannosidase	PREDICTED: serine/arginine repetitive matrix protein 2 isoform 6 [Dasypus novemcinctus]	29.14	singleton	functional approach
3248446	Neopullulanase	hypothetical protein EDEG_03388 [Edhazardia aedis USNM 41457]	29.73	singleton	functional approach
3470944	Cyclodextrinase	hypothetical protein [Clostridium leptum]	29.91	singleton	functional approach
3275619	Alpha-glucosidase	chromosome segregation protein SMC [Thermotoga lettingae TMO]	29.93	singleton	functional approach

b)

3347995	Alpha Amylase	glycosyl hydrolase, family 57 [Sulfurihydrogenibium azorense Az-Fu1]	74.09	singleton	functional approach
2218834	Alpha Amylase	glycosidase [Fervidobacterium pennivorans DSM 9078]	74.29	assembled gene	functional approach
2987017	Pullulanase	Neopullulanase [Caloramator australicus]	74.55	assembled gene	functional approach
4323904	Alpha-mannosidase	PTS system, fructose subfamily, IIC subunit [Thermoanaerobacter italicus Ab9]	74.63	region extracted from assembled contig	conservative approach
3297814	Alpha Amylase	alpha amylase catalytic subunit [Fervidobacterium nodosum Rt17-B1]	74.64	singleton	functional approach

c)

3113705	Alpha Amylase	alpha-amylase [Anoxybacillus sp. SK3-4]	97.86	singleton	functional approach
3290933	Pullulanase	pullulanase [Anoxybacillus sp. SK3-4]	98.67	singleton	functional approach
3333134	Neopullulanase	pullulanase [Thermus igniterrae]	100.00	singleton	functional approach
2551390	Glucoamylase	Glucoamylase-like protein [Thermoanaerobacter thermohydrosulfuricus]	100.00	assembled gene	functional approach
3342434	Alpha-glucosidase	copper amine oxidase [Thermoanaerobacter indiensis]	100.00	singleton	functional approach

Fig. 2. Use of BLAST (2) to screen the protein sequences of potential carbohydrate-utilising enzymes. The 568 hits identified using the specific HMM profiles from metagenome libraries enriched on β-glucans were used for BLAST analysis. The best hits in each case were filtered according to the degree of sequence identity: a) five low identity hits, b) five medium identity hits, c) five high identity hits

The Solr search engine is integrated into the database to allow intelligent text search based on the annotations associated with the KOs and the CAZy database. This also finds sequences whose annotations are not the typed search terms, but may be related. This is illustrated by the results of a search with the term 'amylase' (Fig. 3). Apart from the expected genes, which had been annotated as various classes of α-amylase, genes that were annotated as solute carrier family 3 were also found. When the BLAST function was invoked for such a gene, the conserved domain search showed that the genes might encode transport proteins belonging to a family that also included carbohydrate transport proteins. This ability to carry out searches, browse the results and download sets of selected sequences is useful in broadening the scope of searches and suggesting ideas for detecting novel enzymes.

Discussion

The MEGGASENSE platform allows the rapid generation of specialised databases for genomes and metagenomes. The use of a data warehouse (Fig. 1) allows the incorporation of several relational databases and different

Fig. 3. Example of using the search system in the AMYLOMICS database (http://mrcina.bioinfo.pbf.hr/Amylomics/)

objects (implemented in ZODB) as well as raw data. This is achieved using common loading programs for different databases. It is also easy to customise the graphical user interface.

The coupling of the core functional analysis to the structure of the KEGG database allows good annotation and the ability to view the hits on KEGG pathway charts, which makes it easier to assess whether complete pathways are likely to be present. BLAST analyses with KEGG orthologues will yield good results if the query organism is closely related to the KO member sequences. For more distantly related organisms, the use of HMM profiles as in the MEGGASENSE core analysis will yield better results. This was important for the ZoophyteBase for the proteome of the coral *Acropora digitifera* (*27*). In this case, the identification of functions together with the assignment of KEGG BRITE categories and the use of the Solr search engine allowed the development of novel hypotheses concerning functions such as interaction with symbiotic organisms.

Metagenomic data require assembly of the reads and functional annotations. In addition to the standard approach of assembly followed by annotation of deduced protein sequences, MEGGASENSE offers an additional 'functional' strategy in which *in silico* translated reads are scanned with the HMM profiles prior to the attempts to assemble the reads giving hits to a particular profile. In the case of metagenomes derived from hot springs in the AMYLOMICS database, this resulted in 35 % more predicted genes in contigs with two or more reads. The performance of this functional strategy will depend on the metagenome and the used assembler. However, a better performance of an assembler program on such pre-sorted smaller datasets is likely to be a general property. The MG-RAST server (*11*) identified far fewer KEGG BRITE genes than the MEGGASENSE-based AMYLOMICS database (Table 1). However, this does not accurately reflect the ability of the MG-RAST system to identify gene function as the KEGG BRITE assignment uses BLAST with stringent parameters. MG-RAST assigns genes to cluster of orthologous group (COG) classes (*28*). COG is in many ways similar to the use of KO (KEGG orthologue) in KEGG (*6*). However, the KEGG BRITE classification in which there is a hierarchical organisation of gene function and the fact that KEGG is a highly curated database makes KEGG convenient for functional studies. It would be possible to incorporate COG analyses in a MEGGASENSE database, *e.g.* by generating HMM profiles from COG sequences. MG-RAST is primarily designed to compare metagenomes and provide an overview for large datasets. Although it can be useful for data mining, this will usually involve downloading sequences for analysis in further programs.

The databases reported here are based on the 454 pyrosequencing technology, which gives relatively long reads of good quality. The Illumina technology (http://www.illumina.com (*29*)) yields much larger numbers of reads, which are shorter. When the functional assembly approach was attempted for a metagenome with short Illumina reads (approx. 100 b), the functional approach did not give any advantages as the sequences were too short to give good hits with the HMM profiles (data not shown). However, such data can be incorporated in MEGGASENSE-generated database using standard assembly programs. The Illumina technology can also produce longer reads (>200 b), which would probably be amenable to functional assembly. Single molecule DNA sequencing methods (single molecule real-time sequencing (SMRT), http://www.pacb.com (*30*) or nanopore sequencing, https://nanoporetech.com/ (*31*)) generate long reads with high error rates. It is likely that functional assembly would provide little advantage for long reads, especially as they are likely to have errors resulting in frame shifts in open reading frames.

The modular nature of the platform makes it easy to incorporate any analyses which are useful for the considered problem. In the case of the AMYLOMICS databases, extra specialised HMM profiles for carbohydrate-modifying enzymes were incorporated. The standard integration of BLAST searches in the database allowed the users to compare sequences with the NCBI nr database (*3*). This helped rejection of sequences that were probably falsely identified as carbohydrate-modifying enzymes. It also allowed identification of hits to particular enzyme classes that are distant in sequence from the known members, which are candidates for enzymes with novel properties. The REDPET database (http://redpet.bioinfo.pbf.hr/REDPET; Gacesa et al, in preparation), which was also generated by MEGGASENSE, is designed for bioprospecting hydrocarbon-degrading enzymes in metagenome sequences from the Adriatic Sea. It uses a similar approach to the AMYLOMICS database incorporating appropriate specialised HMM profiles. However, it would also be possible for MEGGASENSE to incorporate more sophisticated analyses tailored to a particular bioprospecting task.

The ever increasing amount of sequence data makes it attractive to use a database-generating platform such as MEGGASENSE to allow more effective exploitation of the data. The ability to incorporate any appropriate form of analysis makes it superior to the use of standard database formats, which are not adapted to the challenges of new projects.

Conclusions

Analyses of metagenomes and genomes are greatly facilitated by incorporating the data in relational databases. Many bioprospecting projects require specialised analyses, which are not provided by standard databases. The MEGGASENSE platform allows the construction of bespoke databases with a core of common search and analysis tools. The use of HMM profiles for functional analysis offers advantages especially for metagenome projects.

Acknowledgements

This work was supported by the European Commission FP7 (265992 to J.Z., S.K.P., O.H.F., R.G., J.D., G.O.H., D.H., A.S.), the Croatian Science Foundation (09/5 to D.H.), the German Academic Exchange Service (DAAD) and the Ministry of Science, Education and Sports, Republic of Croatia (cooperation grant to D.H. and J.C.), and King's College London, UK (to P.F.L.).

Conflict of interest

The presented pipeline was created by SemGen Ltd. as part of FP7-funded project 'Amylomics'. The authors Ranko Gacesa, Jurica Zucko, Janko Diminic, Daslav Hranueli and Antonio Starcevic are employed by SemGen Ltd, which can be contracted to provide third party services. The publication of this paper will serve as an advertisement for some of these services. The authors Ranko Gacesa, Jurica Zucko, Janko Diminic, Daslav Hranueli and Antonio Starcevic declare conflict of interest.

References

1. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res. 2012;40:e9. https://doi.org/10.1093/nar/gkr1067
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10. https://doi.org/10.1016/S0022-2836(0J)80360-2
3. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2015;43:D30–5. https://doi.org/10.1093/nar/gku1216
4. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33:5691–702. https://doi.org/10.1093/nar/gki866
5. Meyer F, Overbeek R, Rodriguez A. FIGfams: yet another set of protein families. Nucleic Acids Res. 2009;37:6643–54. https://doi.org/10.1093/nar/gkp698
6. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109–14. https://doi.org/10.1093/nar/gkr988
7. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23:205–11. https://doi.org/10.1142/9781848165632_0019
8. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2016;42:D279–85. https://doi.org/10.1093/nar/gkv1344
9. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. 2009;37:D233–8. https://doi.org/10.1093/nar/gkn663
10. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods. 2009;6:673–6. https://doi.org/10.1038/nmeth.1358
11. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008;9:386. https://doi.org/10.1186/1471-2105-9-386
12. ZODB – a native object database for Python. Richardson, TX, USA: Zope Foundation Inc.; 2013. Available from: http://www.zodb.org/.
13. Enterprise search platform Lucene/Solr. Wakefield, MA, USA: The Apache Software Foundation; 2013. Available from: http://lucene.apache.org/solr/.
14. Tomcat web server. Wakefield, MA, USA: The Apache Software Foundation; 2013. Available from: http://tomcat.apache.org/.
15. HTML – the language for building web pages; 2017. Available from: http://www.w3schools.com/default.asp.
16. JavaServer Pages (JSP) tutorial; 2017. Available from: http://www.jsptut.com/.
17. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. https://doi.org/10.1186/1471-2105-10-421
18. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195
19. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a web browser. BMC Bioinformatics. 2011;12:385. https://doi.org/10.1186/1471-2105-12-385
20. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003;31: 3497–500. https://doi.org/10.1093/nar/gkg500
21. Yin Y, Mao X, Yang JC, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012;40:W445–51. https://doi.org/10.1093/nar/gks479
22. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009;25:1335–7. https://doi.org/10.1093/bioinformatics/btp157
23. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, et al. Rfam: Wikipedia, clans and the 'decimal' release. Nucleic Acids Res. 2011;39:D141–5. https://doi.org/10.1093/nar/gkq1129
24. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95:315–27. https://doi.org/10.1016/j.ygeno.2010.03.001
25. Kelley DR, Salzberg SL. Clustering metagenomic sequences with interpolated Markov models. BMC Bioinformatics. 2010; 11:544. https://doi.org/10.1186/1471-2105-11-544
26. Barbe V, Bouzon M, Mangenot S, Badet B, Poulain J, Segurens B, et al. Complete genome sequence of Streptomyces cattleya NRRL 8057, a producer of antibiotics and fluorometabolites. J Bacteriol. 2011;193:5055–6. https://doi.org/10.1128/JB.05583-11
27. Dunlap WC, Starcevic A, Baranasic D, Diminic J, Zucko J, Gacesa R, van Oppen MJ, Hranueli D, Cullum J, Long PF. KEGG orthology-based annotation of the predicted proteome of Acropora digitifera: ZoophyteBase – an open access and searchable database of a coral genome. BMC Genomics. 2013;14:509. https://doi.org/10.1186/1471-2164-14-509
28. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28:33–6. https://doi.org/10.1093/nar/28.1.33
29. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9. https://doi.org/10.1038/nature07517
30. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods. 2013;10:563–9. https://doi.org/10.1038/nmeth.2474
31. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17:239. https://doi.org/10.1186/s13059-016-1103-0