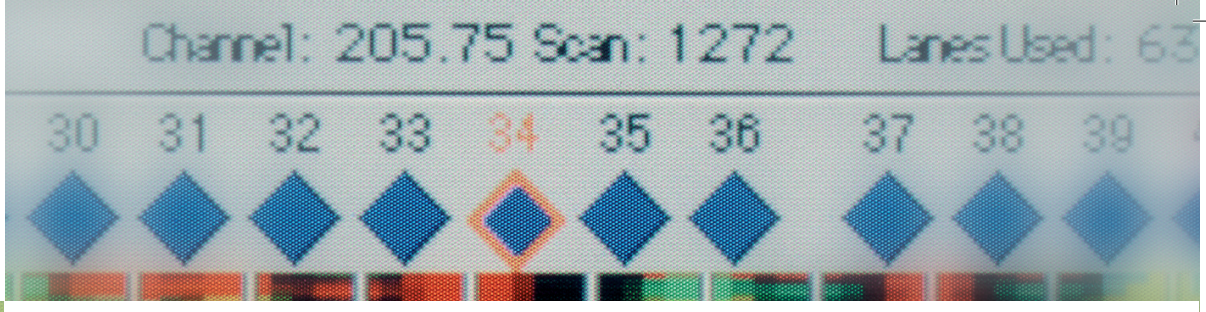


The potential of applying bioinformatics approaches in the risk assessment of genes that lack a function annotation

CGM 2012-03

ONDERZOEKSRAPPORT

The potential of applying bioinformatics approaches in the risk assessment of genes that lack a function annotation

Mogelijkheden en Onmogelijkheden van de bio-informatica: op weg naar een beoordelingsmethodologie voor genen met een onbekende functie

2012
COGEM Research Report CGM 2012-03

author: Christof Francke
affiliations: Center for Molecular and Biomolecular Informatics, NCMLS,
Radboud University Nijmegen Medical Center, Nijmegen;
Kluyver Center for Genomics of Industrial Fermentation, Delft;
Netherlands Bioinformatics Center, Nijmegen;
TI Food and Nutrition, Wageningen.

This report was commissioned by COGEM and BGGO. The content of this publication is the sole responsibility of the author and does not necessarily reflect the views of COGEM and BGGO.

Dit rapport is in opdracht van de Commissie Genetische Modificatie (COGEM) en het Bureau Genetisch Gemodificeerde Organismen (BGGO) samengesteld. De mening die in het rapport wordt weergegeven is die van de auteur en weerspiegelt niet noodzakelijkerwijs de mening van de COGEM of BGGO.

De uitvoering van het onderzoek is begeleid door een commissie samengesteld uit de volgende leden:

Prof. Dr. P. W. M. Hermans (voorzitter)	COGEM; Moleculaire Infectiologie, Universitair Medisch Centrum St. Radboud
Dr. J. Kooter	COGEM; Epigenetica, Vrije Universiteit Amsterdam
Prof. Dr. J. T. den Dunnen	Leiden Genome Technology Center, Leids Universitair Medisch Centrum
Dr. J. E. N. Bergmans	Bureau GGO
Dr. D. C. M. Glandorf	Bureau GGO
Dr. C. J. B. van der Vlugt	Bureau GGO
A. T. A. Box, B. Asc	COGEM-secretariaat

VOORWOORD

Voor het uitvoeren van een risicoanalyse voor werkzaamheden met genetisch gemodificeerde organismen (ggo's) binnen laboratoria, is in de 'Regeling ggo' een beoordelingsmethodologie vastgelegd. Deze methodologie helpt zowel de aanvrager als de vergunningverlener bij het vaststellen van het juiste inperkingsniveau voor handelingen met ggo's. Uitgangspunt bij de beoordeling van de risico's van werkzaamheden met ggo's is kennis over de functie van de te kloneren sequentie, bijvoorbeeld op basis van gegevens over de functie in het organisme waar de sequentie uit afkomstig is.

Nieuwe ontwikkelingen in de moleculaire biologie, met name binnen de werkvelden van genomics en synthetische biologie, leiden ertoe dat er steeds vaker sequenties toegepast kunnen gaan worden waarvan de functie in eerste instantie op puur theoretische gronden is afgeleid. Dit betekent dat er een nieuwe beoordelingsmethodologie ontwikkeld moet worden voor te kloneren sequenties waarvan de functie vooraf niet proefondervindelijk is bepaald.

Binnen de bioïnformatica zijn vele programma's beschikbaar die o.a. op basis van sequentie-informatie, domeinstructuren en fylogenie een voorspelling kunnen doen over de functie van een bepaalde (coderende) sequentie.

Het in dit rapport beschreven onderzoeksproject heeft beoogd te onderzoeken in hoeverre de bioïnformatica in staat is een voorspelling te doen over de mogelijke functie van een (coderende) sequentie. Het project heeft zich daarbij gericht tot sequentiegerelateerde vergelijkende benaderingen en met name aandacht besteed aan het in kaart brengen welke technieken binnen de bioïnformatica een helpende hand kunnen bieden, en wat de mogelijkheden en beperkingen zijn van deze technieken.

Prof. Dr. P. W. M. Hermans
Voorzitter begeleidingscommissie

VERANTWOORDING en DANK

Dit rapport is geschreven naar aanleiding van een vraag vanuit de Commissie Genetische Modificatie (COGEM) en het Bureau GGO, omtrent de mogelijke toepassing van bioinformatica ten behoeve van de functievoorspelling van nieuw geconstrueerde genen. Na een eerste overleg met de begeleidingscommissie is er vanwege de beschikbare tijd en expertise voor gekozen de beschrijving te beperken tot sequentiegerelateerde vergelijkende benaderingen voor de analyse van mogelijke effecten geïnduceerd door modificaties aan genen. De keuze is gebaseerd op twee praktische argumenten. In de eerste plaats sluit de benaderingswijze naadloos aan op de huidige praktijk van de milieurisico-beoordeling: Het mogelijke effect van een modificatie wordt van geval tot geval bestudeerd en de inschatting van het met de modificatie samenhangende risico wordt gedaan op grond van de aanwezige achtergrondkennis over de originele sequentie en over de beoogde gastheer. In de tweede plaats is de vergelijkende sequentie analyse het verst ontwikkeld in termen van het gestandaardiseerd zijn van methoden en algoritmen. Bovendien zijn de methoden en algoritmen veelal via web applicatie beschikbaar en dus ook te gebruiken zonder kennis van programmeren.

Naast een sequentiegerelateerde benadering zou ook een meer data- gedreven benadering in aanmerking kunnen komen ter versterking van de risico analyse. Echter, alhoewel een datagedreven benadering de overhand heeft in het huidige biologisch onderzoek dan is dat wel voornamelijk in verkennende zin. In de bioinformatische analyse van zogenaamde ‘high throughput’ data wordt uitspraak gedaan over de correlatie tussen de aanwezigheid van bepaalde genen/eiwitten en bepaalde fysiologische omstandigheden. Dientengevolge is een datagedreven benadering vooralsnog minder geschikt voor toepassing in een risico analyse op voorhand. Daarentegen, wanneer een vermoeden bestaat van een mogelijk nadelig effect gelieerd met bepaalde genen/eiwitten, dan kan de analyse van ‘high throughput’ data een krachtige manier vormen om risico’s van modificaties in te schatten.

Ik heb getracht in het rapport vooral de gedachte achter de aanpak, en de weging van tools en sequentiegerelateerde informatie, te verduidelijken. Een zinvolle interpretatie van de gegevens verkregen met behulp van bioinformatische technieken is in mijn ogen onmogelijk zonder begrip van de achterliggende gedachte. Het toepassen van de sequentiegerelateerde technieken heb ik vervolgens meer schetsmatig beschreven en geïllustreerd. Ik realiseer mij dat deze schetsmatigheid vaak onvoldoende basis is om er direct een recept of protocol uit af te kunnen leiden. De precieze invulling van het recept of protocol zal van geval tot geval verschillen en vergt in de eerste plaats inzicht in de werkwijze.

Ter illustratie van de verschillende aspecten van sequentiegebaseerd bioinformatisch onderzoek heb ik voornamelijk geput uit mijn eigen onderzoekspraktijk, namelijk de vergelijkende analyse van het bacterieel metabolisme en de daaraan gerelateerde regulatie. De gebruikte voorbeelden zijn niet in alle gevallen direct te vertalen naar de andere biotechnologische werkpaarden, gisten/schimmels en planten, maar in de meeste gevallen wel. Het belangrijkste verschil tussen prokaryoten en eukaryoten is, in die zin, de grootte van het genoom en de afstand tussen de genen. In prokaryoten omvat de gen context vaak ook een ander gen(en) en de aan dat gen(en) gerelateerde functie

informatie kan worden gebruikt bij de interpretatie van de functie van het gekozen gen. In eukaryoten omvat de gen context 'slechts' informatie over regulatie. Maar die informatie is wel degelijk goed te gebruiken bij de interpretatie van de functie van een gekozen gen. Daarbij blijven, ondanks het verschil in de organisatie van het genoom, de principes van de sequentie analyse onveranderd. Het laatste aspect was reden te meer om vooral de gedachte achter de analyse aandacht te geven. Ik hoop dat het rapport in zijn opzet is geslaagd.

Tot slot wil ik hierbij de begeleidingscommissie hartelijk bedanken voor het constructief meedenken over richting en inhoud, en voor de plezierige samenwerking.

Christof Francke, Maart 2012

SAMENVATTING

De huidige strategie voor de risico analyse van een genetische modificatie is gebaseerd op een inschatting van de effecten van de modificatie op grond van de aanwezige kennis over een niet-gemodificeerde tegenhanger. De huidige ontwikkelingen in de biotechnologische toepassing van genetische modificatie vormen daarbij een uitdaging, niet alleen in termen van de schaal waarop modificaties kunnen worden geïntroduceerd, maar ook in termen van wat kan worden gezien als functionele tegenhanger bijvoorbeeld bij het introduceren van vreemde (niet-eigen) genen. Dit rapport beschrijft de mogelijkheden tot het gebruik van sequentie-gebaseerde bioinformatica ter identificatie van de functie van zulke vreemde genen. Daarbij ligt de nadruk op de achterliggende concepten en de praktische toepassing daarvan. Het gebruik van evolutionaire conservering als argument om de functiegelijkheid van twee sequenties vast te stellen wordt beargumenteerd. Daarnaast wordt het IT-gereedschap besproken dat kan worden gebruikt om sequenties en hun functie te analyseren. De toepassing van de concepten en het gereedschap bij de annotatie van genen en regulatoire elementen wordt vervolgens geïllustreerd. Wij concluderen dat een bioinformatische benadering effectief kan zijn om sequenties met gelijksoortige functie te identificeren, maar dat de daarbij te volgen procedure niet zonder meer te standaardiseren is en identificatie daarom vaak aanzienlijke ervaring vereist. Bovendien is een juiste interpretatie van gevaar en/of risico voor een groot deel afhankelijk van de beschikbaarheid van betrouwbare referentie data. Tegelijkertijd is er ten behoeve van een snelle interpretatie duidelijk behoefte aan computerprogramma's die de sequentiegerelateerde data op een overzichtelijke manier samenbrengen.

SUMMARY

Strategies for the environmental risk assessment of a particular genetic modification rely for the largest part on the interpretation of the expected effects induced by that modification in light of the available knowledge concerning an unmodified counterpart. With the advance in biotechnological applications this strategy is challenged, both in terms of the scale at which modifications can be introduced and in terms of the identifiability of the appropriate counterpart. The latter is for instance related to the extent of the modification or to the introduction of unknown sequences. This report explores the potential use of sequence-based bioinformatics approaches to assert the function of new (putative) genes. The main focus in the description is on the conceptual and practical background of comparative sequence analysis. The report therefore includes a discussion of how evolutionary conservation can be used as a signal to infer functional equivalency between sequences, besides a discussion of the available tools and pipelines to search, align and analyze sequences and their function. The application of the concepts and tools is illustrated for the function annotation of several genes and regulatory elements. We conclude that bioinformatics approaches can be of great help in the identification of sequences with similar function, but that this is not achieved in a standard fashion just like that and often will require substantial expertise. Moreover, the appropriate interpretation of the retrieved gene-functions in terms of hazard and/or risk is not straightforward and will, among other things, greatly depend on the availability of reliable reference data(bases). At the same time, we signal a clear need for tools that present the data related to the comparative analysis in an inte-grated way to enable a concise interpretation.

1) CHALLENGES IN GMO RISK ASSESSMENT IN THE GENOMIC ERA AND THE NEED FOR BIOINFORMATICS

a) Introduction

The advent of high throughput genomics technology has spurred the diversity in the biotechnological application of genetic modification in terms of scale and in terms of directed-ness (Carr and Church 2009). The increased availability of genome sequences has allowed a deeper insight in the organization of the genetic material, the variability of that material, and the genetic similarity and difference between organisms. The traditional approaches in the field of microbial biotechnology have profited from the expansion in knowledge and technical capabilities in the sense that difficult constructs (e.g. chimeric genes or genes with multiple mutations) can be made and tested more easily. The same holds in the fields of floriculture (Chandler and Tanaka 2010; Chandler and Brugliera 2011) and crop development (Varshney et al. 2009; Edwards and Batley 2010; Mochida and Shinozaki 2010). Simultaneously, the increase in the amount of data and derived knowledge opens up new roads. For instance in the engineering of microorganisms or plants, knowledge of the system, i.e. the availability of gene-based reconstructions and models, enables optimizing the design, and the high throughput methods enables the rapid screening of a large number of designs (Danchin 2004; Nielsen and Jewett 2008; Mochida and Shinozaki 2011). The large scale also allows for evaluating the properties of randomly generated gene/protein sequences using microorganism as a micro-laboratory (Fisher et al. 2011). In the field of synthetic biology the accumulated knowledge on gene function and regulation, and the improved DNA and protein synthesis capabilities (Ma, Saaem, and Tian 2011), are being exploited to rationally design genetic modules *de novo.* These modules are used for applications in biomedicine (Weber and Fussenegger 2011), the improvement of crop (Gaeta et al. 2011), or the synthesis of chemicals using micro-organisms as factories (Bayer 2010; Boyle and Silver 2011; Ellis and Goodacre 2011).

These developments pose new challenges to the assessment of risk with respect to the application of genetic modification. One of the main challenges is that one has to cope with an ever increasing amount of, and variability in, the data that has to be judged. The current practice is comparative by nature and involves an assessment on a case-by-case basis, so that the potential effect of every single sequence (change) has to be evaluated in light of what is known for the original sequence and the organism that is used. The procedure becomes complicated the moment the number of changes is significant and it becomes almost unworkable in view of the time needed, the moment the potential effect of hundreds or thousands of random sequences are to be analyzed in depth. Bioinformatic approaches might alleviate some of the problems related to the large quantity of data. Moreover, they could add relevant information to the analysis by identifying formerly not recognized links between sequence and function information.

In view of the above COGEM (Commissie Genetische Modificatie) and the RIVM/GMO Office have formulated the following research questions: “Can bioinformatics techniques and approaches contribute to the determination of the function of a particular sequence or to the exclusion of harmful functions?” and “What are the limitations connected to the related techniques and approaches?”. The main aim of the ensuing research is to harness the decision process such that the new challenges can be coped with in a more comprehensive way. From a practical point of view the questions can be reformulated as: “How can one go about sequence-based function annotation?”, “What tools and information should be used?” and “How reliable is the outcome?”. In this report we will give a background to the current risk assessment procedure and how bioinformatics approaches could fit in (chapter 1); we will describe the process of comparative sequence-based function annotation, the tools and the underlying concepts (chapter 2); we will illustrate the potential application of the concepts and tools in the analysis of protein or DNA/RNA sequences (chapter 3); and give some conclusions and future perspectives (chapter 4).

b) Current practice and procedures

There is global consensus about the principles behind a GMO environmental risk assessment procedure. Nevertheless, the associated legislation and procedures vary considerably between different countries (Paoletti et al. 2008). In essence, the risk of a certain modification is derived based on an estimate of the difference in behavior that is induced between the unmodified counterpart and a modified entity. In Europe the principles of the procedure have been laid down in several guidance documents by the European Food Safety Authority (EFSA), each directed at specific groups of organisms like microorganisms (EFSA 2011), plants and crops (EFSA 2006), or animals (Henry et al. 2009). In addition, a clear distinction is made in the kind of application, like different forms of contained use (e.g. in laboratory research or industry) or the deliberate release into the environment. It is required that the assessment is structured and involves, besides an evaluation of intended and non-intended differences, in the first place an identification of potential hazards (Sparrow 2010). The potential risk of a certain GMO application is then determined on basis of the potential hazard and an estimate of the chance that the related properties are evoked during use. There are numerous papers suggesting improvements to the procedure like using preliminary data (Bergmans et al. 2008), the application of other frameworks (Kuiper and Davies 2010), or the explication of test hypotheses (Johnson et al. 2007; Romeis et al. 2008). There is also criticism on for instance the lack of transparency related to the factual basis of the assumptions made in some assessments (e.g. (Tamis, van Dommelen, and de Snoo 2009)). The overall procedure is hardly disputed nonetheless. Currently, the biggest challenge for the improvement and development of suitable GMO risk assessment strategies is the before-mentioned expansion in the diversity of biotechnological applications, or areas of application. For example, guidelines have to be fused to handle the production of pharmaceuticals using GMO plants (Spok et al. 2008), whereas new guidelines were drafted for commercial parties synthesizing DNA or protein sequences (see discussion in (Eisenstein 2010; Erickson, Singh, and Winters 2011). In the latter case, the new US government guideline (US government 2010) stipulates the use of a bioinformatics screening of sequences. And a clear-cut bioinformatic implementation of the screening has already been suggested (Adam et al. 2011). At the same time, experts have commented that manual inspection of the screening results should remain an important part of the overall assessment (Eisenstein 2010).

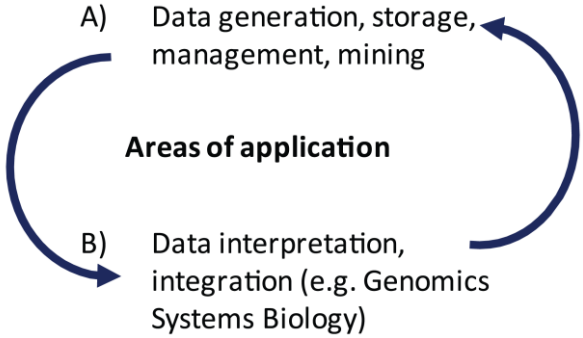


Figure 1. The two main perspectives within the application of bioinformatics in biology.

c) Bioinformatics is a scientific discipline

With the increase in computer power and the almost simultaneous increase in the amount of available biological data, Bioinformatics has emerged as an independent scientific discipline within the biosciences. The discipline has grown rapidly to include a variety of fields that each requires its own specialism, much like in the case of the disciplines Biochemistry and Biophysics. Basically, the discipline Bioinformatics involves the application of information technology and its concepts to the study of biology and medicine. It centers around two main interdependent focus areas which might be summarized as 'data handling' and 'data interpretation' (see Figure 1). The former area includes activities like literature mining, next generation sequencing or mass spectrometry, and the storage and appropriate assembly of the generated data. The latter area includes activities like network reconstruction and modeling, or function annotation based on sequence and structure. Thereby, the use of the appropriate bioinformatics approach is an essential step in the extraction of new knowledge from any combination of high throughput data. The approach can either be

data-driven, i.e. linking available data or information on basis of observed correlations, or knowledge-based, i.e. linking available data/information on basis of biological concepts. A knowledge-based approach is the most applicable in the assessment of the potential consequences of a particular genetic modification. It is currently better developed and relates directly to a mechanistic description of biological phenomena, which makes it better suited for a (more) quantitative assessment of risk.

d) Justification of focus of study

The guiding principle of a knowledge-based bioinformatics approach is that functional information is gathered/connected on the basis of sequence comparisons. Basically, a sequence under study (the query) is used to find similar sequences (the subjects) that are connected somehow to experimental data and/or literature. In case the similarity between the query and a subject is high it can be inferred that the information linked to the subject can also be linked to the query. This procedure fits seamlessly to the initial phase of the current risk assessment procedure, where the knowledge on a subject (e.g. the organism that is used, the original sequence, or the cellular physiology) is used to assess the potential effect of the changes present in the query. Thus a bioinformatics approach may add considerably to the initial phase of the risk assessment that is the identification of potential hazard (Adam et al. 2011; EFSA 2011). In contrast, as current bioinformatics techniques deal primarily with linking/transferring information they are not well-suited to determine risk yet. For the calculation of risk, background models are needed that incorporate the probabilities related to certain events. Although comparative genome analyses may shed light on these probabilities in the future, the related methodology is still far from a general practical application. Similarly, although in principle it should be possible to base a function prediction on sequence and first principles alone, much of the essential knowledge to do this is still insufficient. For instance in the area of protein structure analysis researchers have been busy for many years to predict structure directly from sequence, as witnessed by the two-yearly CASP competition that was held for the 9th time in 2010 (Moult et al. 2011). In recent years the predictions have come closer to reality. Nevertheless, without prior knowledge these efforts are still far from delivering specific function predictions. In other areas advance has been made too, for instance with respect to the sequence-based prediction of protein sub-cellular location (Zhou et al. 2008; Rastogi and Rost 2010). Similarly, the application of bioinformatics techniques in the field of systems biology has led to relatively reliable functional predictions on the basis of gene content alone (Price, Reed, and Palsson 2004; Teusink, Westerhoff, and Bruggeman 2010; Bordbar and Palsson 2011; Orth et al. 2011). But also in the latter case the methodology is not yet applicable without investing a considerable amount of time and expertise. We will limit our description of potentially relevant bioinformatics approaches therefore mainly to those approaches that are easily accessible and which aim primarily to infer functional equivalence on the basis of sequence similarity.

2) PRINCIPLES AND TOOLS BEHIND THE SEQUENCE-BASED TRANSFER OF INFORMATION

A direct way to evaluate a prediction of the encoded function of a DNA sequence is by sequence comparison. Likewise the effect of sequence changes can be evaluated by comparison of the sequence with other sequences of known function. In case the similarity between a pair of sequences is sufficient it can be inferred that they share similar properties, like encode the same protein. During the process of comparison and inference at least three essential choices have to be made: i) that of the sequences that are to be compared (e.g. via the particular choice of a reference database, the comparison of full length or partial length, the use of protein sequence or DNA sequence); ii) that of a similarity measure (e.g. E-value, evolutionary relatedness); and finally iii) that of the level of similarity that can be considered sufficient to decide that sequences probably encode the same function (e.g. by using E-value cut-offs or evolutionary criteria).

In the following sections the principles of sequence comparison and analysis, *casu quo* function annotation will be discussed. First (I), the evolutionary premiss behind comparative sequence analysis will be described together with the associated terminology. Then, the hierarchical nature of function and its consequences for sequence analysis will be discussed; Second (II), the knowledgebase and some issues of data privacy will be described; Third (III), the tools to perform sequence similarity searches, to make multiple sequence alignments or to cluster sequences will be described; And finally (IV), different ways to evaluate what is considered sufficient similarity to allow inference of function will be discussed.

I) CONCEPTS OF EVOLUTION AND FUNCTION

a) The implication of the unity in biochemistry on evolutionary conservation

In 1926 Albert Jan Kluyver and his student Hendrick Jean Louis Donker published their seminal work on “Die Einheit der Biochemie” (Kluyver and Donker 1926) (see also (Kluyver, 1924)). The idea of a uniform biochemistry for all organisms was readily accepted, expanded (e.g. (Monod and Jacob 1961)) and ultimately became one of the cornerstones of comparative molecular biology (Friedmann 2004). The uniformity of chemistry implicitly requires that the proteins involved in cellular metabolism should share similar properties in different organisms. From an evolutionary perspective the necessary similarity is ensured by genetic conservation. Hence the lineage of sequences between species is the main determinant to decide about functional similarity in comparative sequence analyses: a gene that has been evolutionary conserved between two species in general will encode a protein of (near-)identical function in both species.

The appropriate definition of the terms used to describe the evolutionary relationship between two sequences has been a subject of hot debate among geneticists and evolutionary biologists (as described in (Abouheif et al. 1997; Egel 2000; Koonin 2005; Fitch 2000). From a practical point of view we prefer a use that was clearly set out by W.M. Fitch[1]. One of the terms that is often used inappropriately and thereby causes much confusion is ‘homology’. In the practical definition by Fitch, ‘homology’ refers to the evolutionary relation between two sequences that derive from one ancestral sequence. Because of time, the composition of the two sequences will have diverged and as a result the sequences will share a certain percentage identity or similarity (not

1 The evolutionary relationship between genes is described using the following terminology (from Fitch, W. M. 1970. Further improvements in the method of testing for evolutionary homology among proteins. J Mol Biol **49**:1-14, Fitch, W. M. 2000. Homology a personal view on some of the problems. Trends Genet **16**:227-231.):

analogy is the relationship between two genes that have descended from unrelated ancestral genes but have converged to acquire similar functionality.

homology is the relationship between two genes that have descended from a common ancestral gene and have diverged on the sequence level.

orthology is the relationship between two homologous genes that originate from the same gene in the most recent common ancestor of the species that are compared.

paralogy is the relationship between two homologous genes that arose from a gene duplication.

homology!). In this scheme, sequences that share similarity and/or have the same molecular function by virtue of convergence (i.e. because they descended from unrelated ancestral sequences) are termed 'analogous'. The evolutionary relation between homologous genes or proteins is reflected by the use of the term 'gene- or protein-families'.

Another relational term that causes confusion is 'orthology'. We have illustrated the appropriate use in Figure 2. Correct usage of the term is important because orthologous sequences can be inferred to have (near-) identical molecular function. Nevertheless, the biological role of orthologous sequences sometimes varies as the role is not only determined by the properties of the sequence itself but also by the environment the sequence is in. For instance, orthologous genes that are expressed under a different condition, might perform differently (e.g. a change in pH might change the specific activity of a protein). Consequently, it can be assumed that paralogous sequences either have very similar yet distinct molecular functions, or have (near-) identical molecular functions but distinct biological roles (see Figure 2).

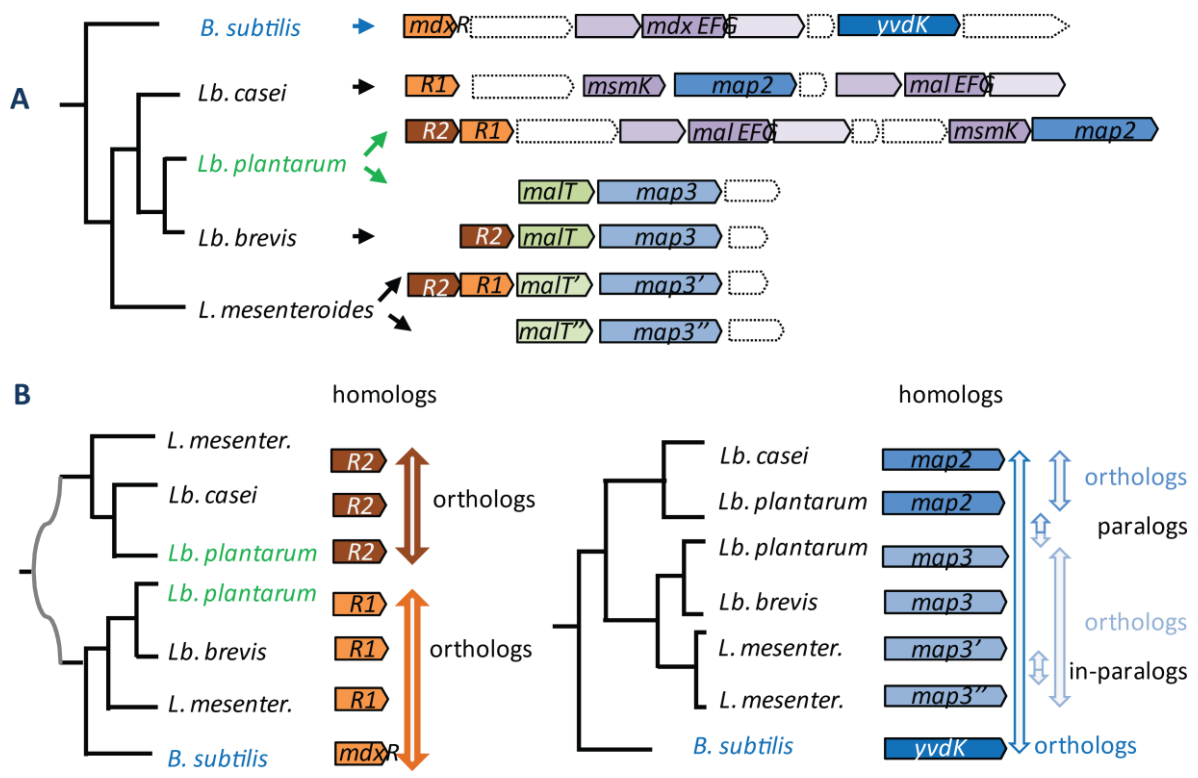


Figure 2. The lineage of the genes related to maltose transport and phosphorylation in *Lactobacillus plantarum*.

A) The phylogenetic relationship between the compared species (left) and the compared gene clusters (right). The displayed clusters are composed of genes that encode a LacI-family transcription regulator (i.e. mdxR, *R1* and *R2*), a maltose-proton symporter of the GPH-family (*malT*) or an oligo-maltose ABC transport system (*malEFG* and *msmK*), and a maltose phosphorylase (*yvdK*, *map2* and *map3*).

B) Description of the evolutionary relationship between some of the genes constituting the clusters. *L. plantarum WCFS1* has 12 genes genes encoding LacI-family transcription regulators (Francke et al. 2008), where *lp_0172* (i.e. *R2*) and *lp_0173* (i.e. *R1*) are more distant family members (i.e. homologs) and are found associated to maltose transport and phosphorylation. *L. plantarum* WCFS1 has 4 genes of the GlycosylHydrolase 65 family (CAZY nomenclature). Map1 and Map3 of *L. plantarum* are paralogs and they were both annotated as a maltose phosphorylase because of the orthologous relationship with Map of *Bacillus subtilis (yvdK)* and because of the conserved gene context. The two *map* homologs found in *L. mesenteroides* are both orthologous to *L. plantarum map3* and as they arose after the last speciation event they are considered in-paralogs.

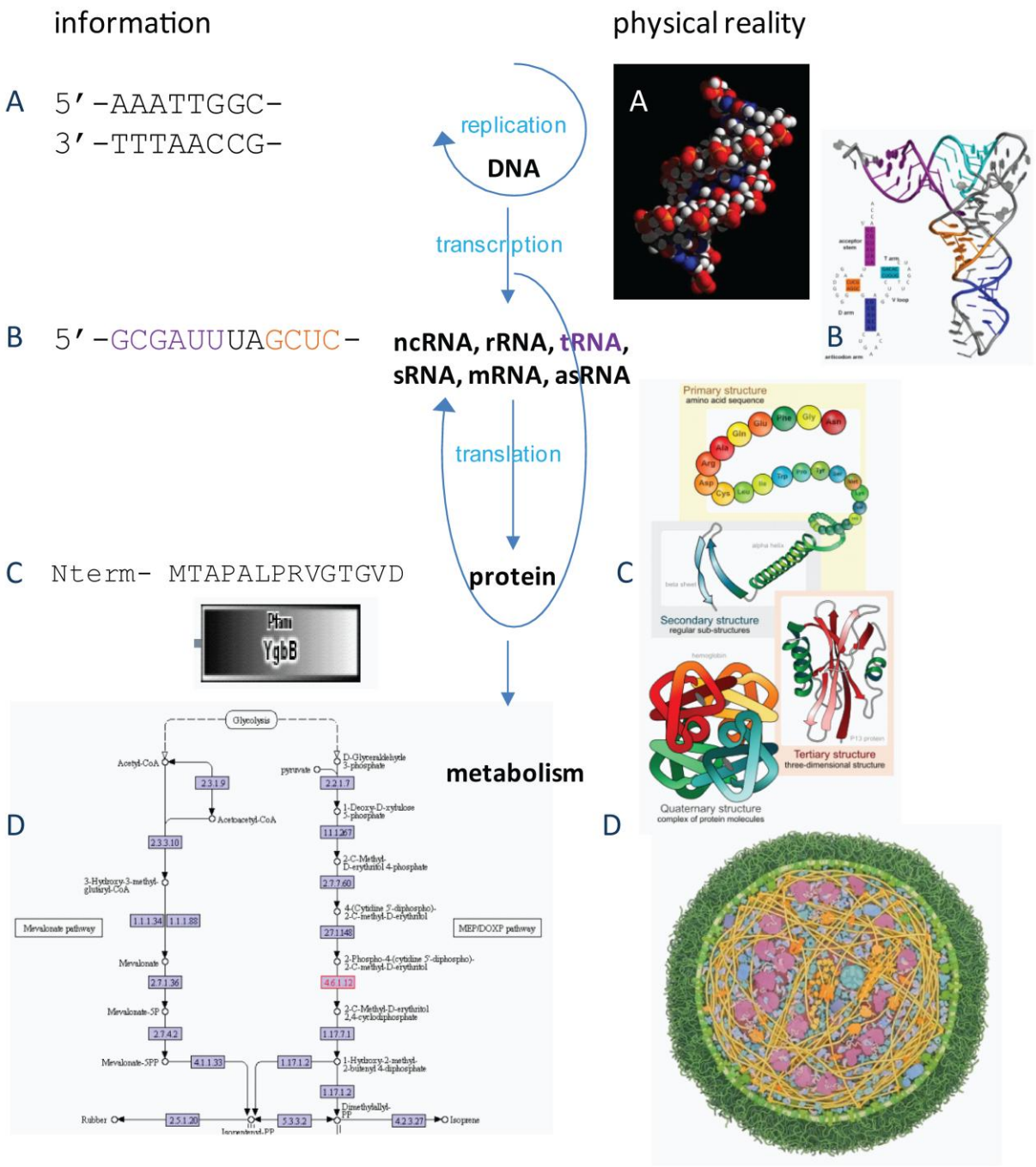


Figure 3. The functional hierarchy within the cell, the flow of information and the molecular nature of life.

Cellular metabolism can be perceived and described at various levels of abstraction. Depicted in the middle is the central paradigm of cellular metabolism: **A**) The DNA represents the genetic information that implicitly encodes all macromolecular players. It is being maintained in progeny by virtue of replication. **B**) The DNA is being transcribed to various forms of RNA, some with regulatory functions while others are involved in the translation to **C)** protein. The proteins catalyze the **D**) metabolic conversions necessary to energize and build up the cell. Depicted on the left is the same hierarchy formulated in terms of molecular information. The partial protein sequence relates to MECDP-synthase (PFAM domain YgbB), which plays a role in terpenoid backbone biosynthesis (KEGG map 00900). Depicted on the right are representations of related molecular structures as they were taken from the wwweb. **A**) courtesy of Christian Ude. **B**) courtesy of Joel L. Sussman et al. JMB, 1978. **C**) courtesy of Mariana Ruiz Villarreal. **D**) The illustration of a mycoplasma cell is a courtesy of David S. Goodsell, the Scripps Research Institute.

A strong indicator of orthology and/or functional equivalency is the conservation of genomic context (Huynen et al. 2000), which can by and large be considered as a conserved co-localization of genes on the genome. In classical genetics, the term 'synteny' was used to describe the physical co-localization of genetic loci on the same chromosome within an individual or a species. In comparative genomics the term is often used to describe the fact that gene-pairs or clusters on the genome have been conserved between species. As a result of these two slightly different definitions, also the use of this term has been subjected to debate (Passarge, Horsthemke, and Farber 1999). The problem is nicely described by Dewey (Dewey 2011). He proposes to use the term 'positional orthologs' or 'toporthologs' to refer to orthologs with conserved genomic context. Nevertheless, the application of toporthology as a concept is not as simple as it might seem. While studying the genomic origin and fate of gene duplicates, it was found that in around 30% of the cases the two most similar gene sequences between two species where not the two sequences with similar context (Notebaart et al. 2005). In the case of these very similar genes the genomic context most probably is the strongest indicator of functional equivalency (Burgetz et al. 2006).

b) A hierarchical perspective on function

Function is a central concept in the process of annotation. However, its use in the description of molecular properties is ambiguous. We will therefore use the term 'function' or 'molecular function' (Bork et al. 1998) to refer to context-independent[2] properties of a molecule (Francke, Siezen, and Teusink 2005). For instance, in the case of a protein the 'molecular function' could be that it can catalyze a certain reaction or has affinity for a specific protein or a specific sequence of DNA. We will use the term 'role' to refer to context-dependent properties of a molecule. For example, a protein does act in a certain pathway, recruits another protein to a protein complex or activates transcription.

2 In the strict sense 'context-independency' does not exist in nature. We refer here to the properties under standard cellular conditions

Another complication in the interpretation of the term 'function' arises from the dichotomy between sequence either viewed as code (i.e. information) or as a molecule (i.e. having a physical structure). Within the discipline of Bioinformatics sequences are treated mostly in terms of the former. Nevertheless, inevitably, the natural constraints act at the molecular level[3]. A final point of consideration is illustrated in Figure 3 and involves the fact that the molecular functions linked to a DNA sequence relate to properties that play a role at different process levels.

In a protein sequence (primary structure) the ordered amino acids form a spatial structure (secondary and tertiary structure) and within this stucture many residues fulfill specific roles, like compose the catalytic site or form a binding surface. Actually, most proteins harbor more functions/roles which are in general not related to the same residues and are also not related to a similar number of residues. Appropriate function annotation is complicated by such multiple functions/roles as one of them in most cases dominates the conserved sequence signature. For example, the ABC multicomponent transport system, which represents one of the largest and most ancient protein superfamilies, is characterized by an ATP-binding protein sub-unit (Dassa and Bouige 2001; Davidson et al. 2008). The capacity to bind and hydrolyze ATP is encoded by an evolutionary conserved set of residues making the functionality easily tractable (e.g. by using sequence profiles). The sub-unit also has a

3 A perfect example of the consequence of this dichotomy can be found in the large variety in lengths of transcription factor (TF) binding-sites predicted on the basis of statistical considerations only (as done in many bioinformatics tools). However, most TFs interact with the DNA by virtue of a protein helix-turn-helix motif (HTH). In molecular terms the HTH slides into the (major-)groove of the DNA. As the DNA is helical the number of exposed nucleotides to the HTH is thereby limited to around 5-7. Therefore the molecular nature of the interaction limits the physical contact between most monomeric TFs and the DNA to 5-7 nucleotides (unless, in special cases, the DNA bends and interacts with other parts of the protein).

second role: binding to the appropriate permease sub-unit. This capacity is encoded by different residues and is much less recognizable as the ABC-permease proteins vary considerably between different substrates. In some cases the binding role has seemingly disappeared as the two sub-units became fused in to one protein. Nevertheless, also in these cases several residues within the ATP-binding domain will be needed to interact with the permease domain such that the liberation of free-energy by the former domain is directly coupled to transport by the latter.

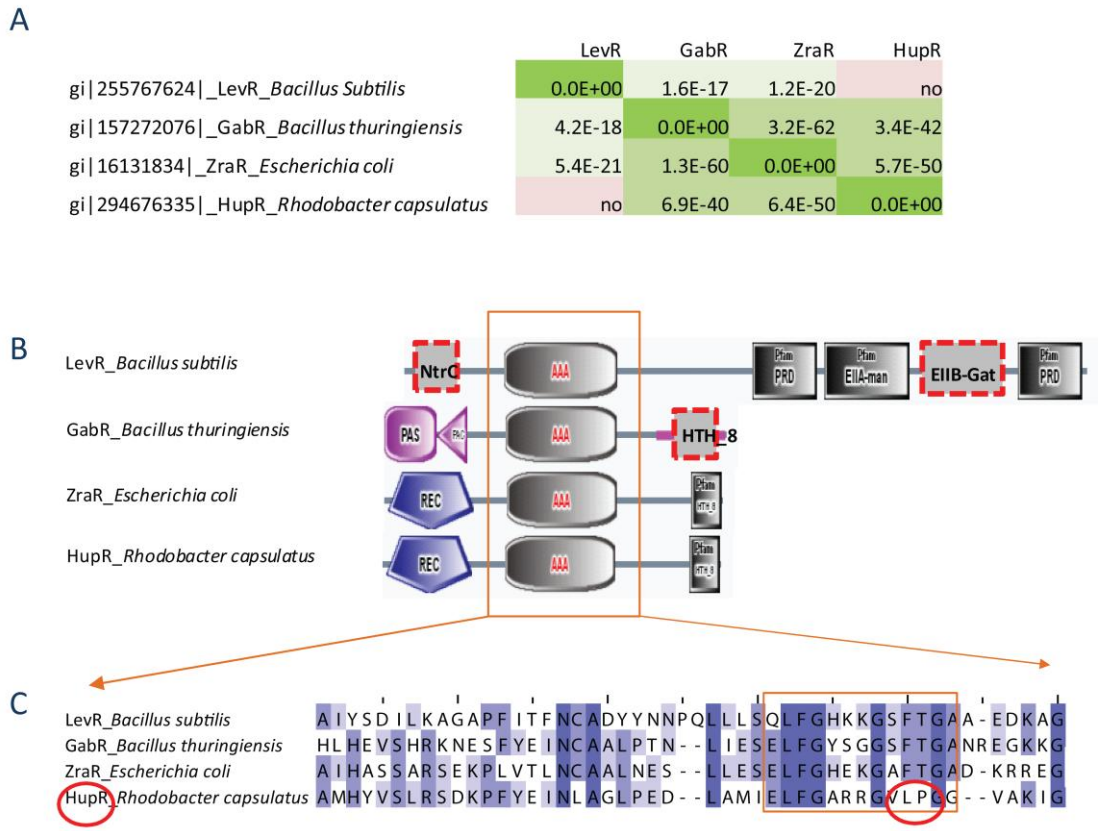


Figure 4. Comparison of the composition and organization of the functional domains of four enhancer binding proteins related to sigma factor 54.

A) BLAST E-values related to the one-to-one sequence comparisons for the Enhancer Binding Proteins (EBPs) LevR of *Bacillus subtilis*, GabR of *Bacillus thuriengiensis*, ZraR of *Escherichia coli* and HupR of Rhodobacter capsulatus. **B)** Domain organization of the four EBPs. The domain composition was determined using SMART (Letunic, Doerks, and Bork 2009). One should realize that the applied Hidden Markov Models in various cases lack the specificity to attribute the functional domain that is present (as indicated by the dashed boxes). C) Multiple sequence alignment of part of the central AAA domain.

In many prokaryotes, sigma factor 54 is a central controller of the transcription of genes related to the biosynthesis of those metabolites that serve as intermediates in the redecoration of the bacterial exterior (Francke et al. 2011). Sigma-54 mediated transcription initiation can only occur after activation by an Enhancer Binding Protein (EBP). There is a considerable variety in EBPs to accommodate responsiveness to a large range of internal and environmental stimuli. The comparison of four different EBPs can be used to illustrate various aspects of protein sequence analysis. Considering the diversity in domain composition, the BLAST E-values that are found are relatively low. This is related to the fact that the central ATPase domain (AAA), which is responsible for the interaction with the sigma factor and for ATP hydrolysis, is common to all EBPs and has a clear sequence signature. Nevertheless, the EBPs can be divided in two main groups on basis of the BLAST E-values. The two groups have a different domain order, where LevR has an N-terminal DNA-binding domain (NtrC) and C-terminal signal recognition domains (PRD, EIIA, EIIB, PRD), and GabR, ZraR and HupR have an N-terminal signal recognition domain (PAS and REC) and a C-terminal DNA-binding domain (HTH8). In the case of HupR the central sequence has diverged and interaction with sigma-54 was lost (substitution FT -> LP (Davies et al. 2009)), which might explain the much higher E-value against LevR. In addition, the divergence of the central domain might also explain the lower E-value for the comparison GabR-ZraR versus ZraR-HupR although the domain composition of the latter two is more similar.

The ATP-binding and hydrolyzing functionality of the domain is also being exploited by other proteins and protein systems. From a perspective of evolution the domain has become genetically associated with these different systems or even has been fused. In fact, the fusion of functional domains is regularly encountered in proteins and in some cases these domains have become shuffled during the course of evolution (Wolf, Wolf, and Koonin 2008; Bornberg-Bauer, Huylmans, and Sikosek 2010). As a consequence, these protein sequences can not be compared directly to derive their similarity, but every domain has to be evaluated on its own (George and Heringa 2002). The multiplicity of functions is illustrated in Figure 4.

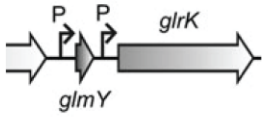
An assessment of the effect of particular changes in a certain sequence, or of the introduction of a certain sequence, will in general be focused on effects at the protein level. This has various reasons. In genetic engineering most of the directed changes are in coding sequence, as they are ultimately aimed to affect the protein activity. Likewise, the introduced changes in random genetic experiments often display the most prominent effect at the protein level. Also in practical terms the protein level is the most accessible for study because of the availability of sufficient functional and structural data. Moreover, the consequence of a change in protein sequence is often much easier to interpret than that of a change in nucleotide sequence because of the higher specificity of the amino acid 'code' (20- vs. 4-letter alphabet). Nevertheless there could be physiological effects that are the result of sequence changes that take effect at a different hierarchical level, like for example in the adaptation of regulatory elements to enable gene expression under varying conditions (see examples in (Hittinger and Carroll 2007; Stoebel et al. 2009)). Unfortunately, these effects are often hard to assess due to a lack of time, data and/or knowledge. For instance, although a large number of regulatory RNAs (Fröhlich and Vogel 2009; Breaker 2011; Guell et al. 2011; Meng et al. 2011; Okamura 2011; Storz, Vogel, and Wassarman 2011) is known and engineering efforts are underway (see e.g. (Culler, Hoff, and Smolke 2010; Liang, Bloom, and Smolke 2011)), much of the variety has not yet been discovered (including many small and non-coding RNAs) and is relatively difficult to assess (as illustrated in Figure 5). Thus, whereas the functional assessment by necessity (in terms of the available time and knowledge) often has to be restricted to effects on the protein level it remains important to be aware that sequence changes might take affect at other cellular process levels.

II) SOURCES OF INFORMATION

In a comparative sequence analysis one of the first difficulties that has to be overcome is the choice of the appropriate resources/databases to search for information. Unfortunately, potentially relevant sequence and function information is found wildly distributed over the world-wide web and in literature. On the one hand there are thousands of databases that focus on biological sequences of particular function, whereas on the other hand the experimental information is still scattered throughout literature because it cannot be easily extracted and indexed. Fortunately, in several areas a consolidation of information has occurred and a search can be limited to a small number of resources/databases. In addition, some institutes have created useful portals that enable a more structured search of the scattered information. In the following we will describe various types of resources/databases that are essential for comparative sequence analyses, give some possibilities to search the literature using sequence information and discuss matters of search privacy.

For the non-informaticists it is important to be aware that most resources use their own characteristic information storage format. As a result the information that can be recovered from multiple resources will strongly depend on the applied identifiers (as illustrated in Figure 6). Moreover, all stored public information should be treated with caution as much of it has been inferred on basis of automated procedures (see (Devos and Valencia 2001; Schnoes et al. 2009)) and/or may not be regularly updated. Therefore, one of the essential activities in the process of comparative analysis of genes is to tract the experimental basis of the function descriptors.

A



B

TTATCTAATAAGTTTATCTTAAGTTGTCTCTTTTTAGCGACACAGTGGCT
GATTTACATTCAAACAGCGGAGGAGACATAAACCCCTCAAAGCAACGTAT
AATCAGTACGTTACCAAACTATTTTCTTTATTGGCACAGTTACTGCATAA
TAGTAACCAGTGGCTCATTCACCGACTTATGTCAGCCCCTTCGGGACGTG
CTACATAAAATACGAATGACGCACAACAAGGTGCCTGCCGTCCAACTTCT
GATATCAGCGTAGCTATATCAACCATCGGGCGAAACGTCGAGTTAGGCAC
CGCCTTATTCCATAACAAAGCCGGGTAATTCCCGGCTTTGTTGTATCTGA
ACTTCCCCTCGGTTAGCATCAGGCTATTCGCGTCTGACGAGAGTAACACC
TTGAAA putative RBS

translation start *glrK*

C

Reference RNA sequences refseq_rna (blastn)

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
XM_002430452.1	Pediculus humanus corporis protein C21orf59, putative, mRNA	37.4	37.4	10%	3.9	100%	G
XM_001635893.1	Nematostella vectensis predicted protein (NEMVEDRAFT_v1g241397)	37.4	37.4	10%	3.9	100%	UG
XM_001634007.1	Nematostella vectensis predicted protein (NEMVEDRAFT_v1g242462)	37.4	37.4	12%	3.9	96%	G
XM_001030881.1	Tetrahymena thermophila ATPase, histidine kinase-, DNA gyrase B-,	37.4	37.4	13%	3.9	92%	GM

Nucleotide collection nr/nt and NCBI genomes (blastn)

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
AP012306.1	Escherichia coli str. K-12 substr. MDS42 DNA, complete genome	331	331	100%	1e-87	100%	
CP002729.1	Escherichia coli UMNK88, complete genome	331	331	100%	1e-87	100%	
AP012030.1	Escherichia coli DH1 (ME8569) DNA, complete genome	331	331	100%	1e-87	100%	
FN649414.1	Escherichia coli ETEC H10407, complete genome	331	331	100%	1e-87	100%	
NC_014228.1	Xenorhabdus nematophila ATCC 19061 chromosome, complete genor	62.6	62.6	60%	2e-06	74%	
NC_012912.1	Dickeya zeae Ech1591 chromosome, complete genome	62.6	62.6	53%	2e-06	76%	
NC_014500.1	Dickeya dadantii 3937 chromosome, complete genome	60.8	60.8	45%	6e-06	77%	
NC_009832.1	Serratia proteamaculans 568 chromosome, complete genome	60.8	60.8	50%	6e-06	75%	

Figure 5. The genomic position of the small RNA regulator GlmY of glucosamine-6-phosphate synthase.

A) GlmY was found encoded upstream of the gene *glrK* (Reichenbach, Gopel, and Gorke 2009), which encodes the Histidine Kinase of a two component system that gauges the stress signals Epi, sulfate, and phosphate (Reading et al. 2009).

B) The regulatory elements connected to the expression of *glmY* and *glrK*. Bioinformatic and experimental analysis of the upstream region of *glrK* in *E. coli* and other Enterobacteriaceae showed that there are two overlapping promoters (sigma 70, brown; and sigma 54, red) preceding *glmY* (bold and underlined) and that expression via the sigma 54 promoter is activated by the two component enhancer binding protein GlrR (=QseF, YfhA) (binding site red underlined) after activation by the two component histidine kinase GlrK (QseE, YfhK) (Gopel et al. 2011).

C) BLASTN search output for GlmY. A search in the NCBI refseq_rna database yielded no hits (lowest E-value was 4 at 10% coverage). A search in the nucleotide collection and genomes database yielded one hit, namely *glmY* in most of the Enterobacteriaceae. However, for the more distantly related genomes like that of Dickeya and Serratia the E-value was already relatively high. This implies that it wil be very hard to find GlmY homologs this way, even if they would be present in a genome.

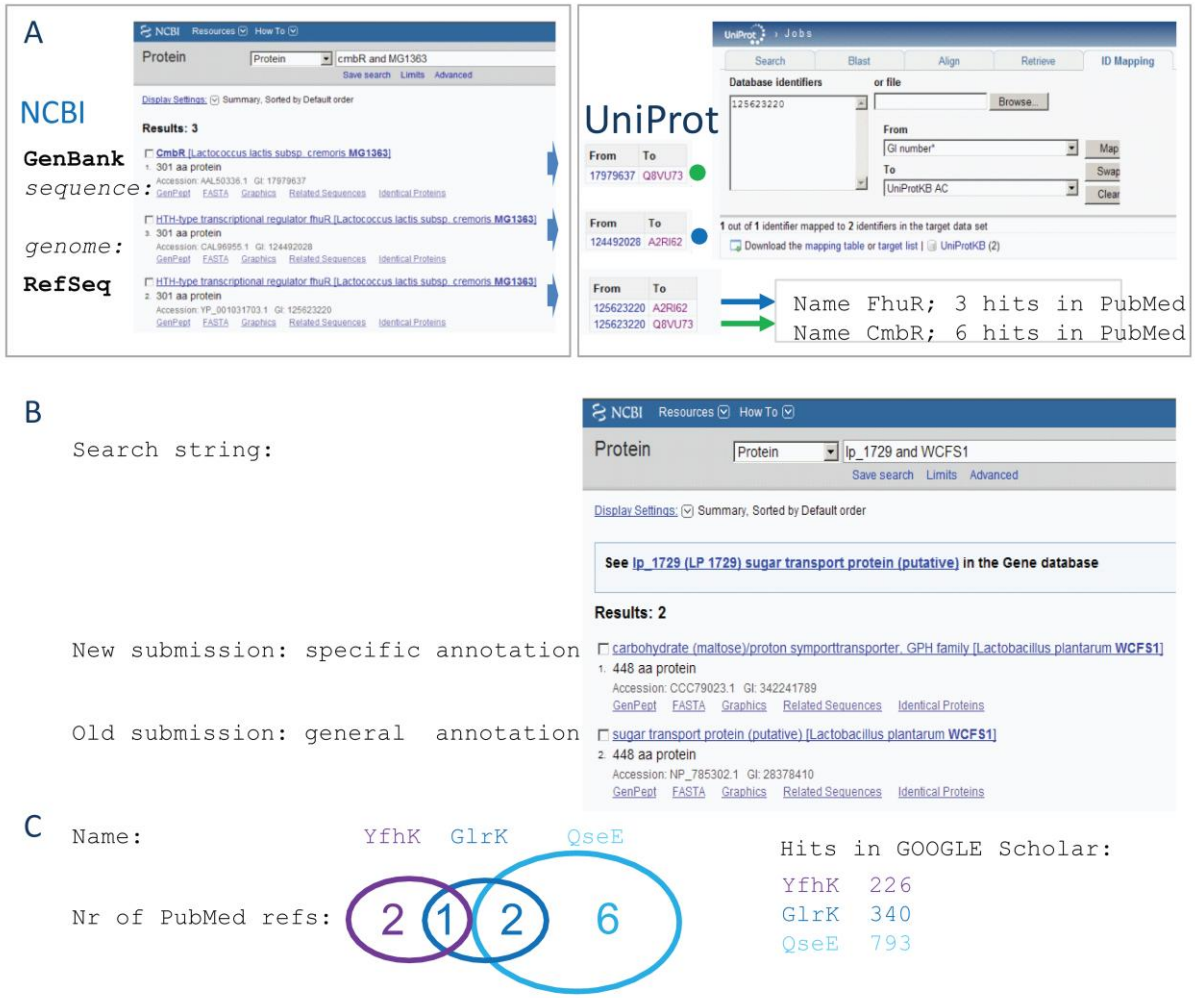


Figure 6. The adequacy of particular sequence identifiers to retrieve relevant data from different databases/resources.

A) Retrieval of gi-codes from the NCBI protein database and gene names from Uniprot for the regulator CmbR=FhuR of *Lactococcus lactis*. **B)** Retrieval of annotation information for the gene *lp_1729* (locus tag) of *Lactobacillus plantarum WCFS1*. The gene is named *malT* in Figure 1. **C)** Number of references retrieved from PubMed and GOOGLE Scholar using the three different protein names related to gene *b2556* (locus tag) of *E. coli K12 MG1655* (see also Figure 5).

Upon submission in the NCBI repository (GenBank) every gene/protein sequence obtains a unique gene identifier (GI) and accession number. For instance in the case of the cysteine-related regulator encoding gene of *Lactococcis lactis*, there have been two individual submissions of the same gene, one using the gene-name *cmbR* and one using the gene name *fhuR* (GI:17979637 and GI:124492028, respectively). At NCBI an additional data-table is created for every gene/protein in the RefSeq database, again with unique sequence identifiers (GI:125623220). Many resources synchronize their databases with that of NCBI. The standard protein data-resource Uniprot also does this and provides a Uniprot-identifier to every sequence with a different name. In this particular case the cysteine-related regulator of *L. lactis* can be found using two Uniprot identifiers, Q8VU73 (GenBank 17979637) and A2RI62 (GenBank 1244920228). Looking up the sequence using the GenBank or Uniprot identifiers will yield only one of the gene names, whereas a search with the RefSeq identifier yields both. Although it might seem a minor point, the used gene name can considerably affect the retrieval of additional data. For instance a pubmed search using the individual gene names and the word 'regulator' yielded 6 papers in the case of '*cmbR*' , whereas only 3 in the case of '*fhuR*'.

a) Public repositories of sequence data

The scientific community is very fortunate that it was soon recognized by US, European and Japanese governmental agencies that the storage of genome and protein sequence information should be centralized. This centralization is enforced by most publishers who require that new sequence data related to a publication are deposited in one of the central resources. For genome sequence information there are three central repositories one @NCBI (USA; (Benson et al. 2012)), one @ EBI (EU; (Leinonen et al. 2011)) and one @ NIG (Japan; (Kaminuma et al. 2011)) (see Table 1). They act as mirrors for each-others information but use their own internal identifier system. This sometimes impairs information retrieval between the resources. For the retrieval of protein sequence information the same resources can be used. However, another logical place to start is the Uniprot knowledgebase (the Uniprot Consortium 2012) maintained at the EBI, SIB (Swiss Institute of Bioinformatics) and Georgetown University. Protein structures and related information can be found in the PDB (Rose et al. 2011).

b) Function classification schemes and public resources of function data

Various kinds of function descriptors have been applied to capture the molecular properties and biological role of nucleotide and protein sequences (Table 2). And most of the deposited gene and protein sequences in the reference databases have been associated to one or more of these descriptors. However, the asserted associations of sequence to function are not necessarily of high quality and regularly incorrect. By the way, this statement is far less true for the manually curated databases like for instance SwissProt. In the case of protein sequences the functional information is often derived from profile searches (see next chapter). The profiles may relate to enzymatic or regulatory function, but also to structural features such as signal peptides or transmembrane helices. One of the most versatile protein profile databases is PFAM (Finn et al. 2010), which also has an RNA-profile counterpart called RFAM (Gardner et al. 2009). In general, the profiles or sequence motifs have been formulated on basis of sequences linked to experimentally validated function and thus provide a link to those data.

Importantly, all of the above-mentioned resources provide the opportunity to search the content using a single query sequence or a number of them, thereby avoiding the use of identifiers that were not 'aligned'. Despite the fact that there has been no clear public incentive to create resources that link specific function information derived from literature to the sequence information present in the general resources, several research groups fortunately felt the need to do so anyway (see Table 3). The largest of these private initiatives is KEGG (Kanehisa et al. 2008), which was developed and is being maintained by the Kanehisha lab in Kyoto, Japan. It provides links between sequence identifiers and chemical information like compounds, reactions and pathways. The information stored within KEGG forms the basis of most other commercial and non-commercial resources of sequence-linked data on function. A second important resource of sequence-linked annotation information is Biocyc, an organism-specific data resource developed and maintained by the group of Peter Karp at SRI International (Caspi et al. 2009). The main representative EcoCyc represents data from over 21000 publications (Keseler et al. 2011). Other pathway databases include PathwayCommons (Cerami et al. 2011) and Wikipathways (Kelder et al. 2009) for mainly animals and yeasts, Reactome (D'Eustachio 2011) for humans, and PMN (Zhang et al. 2010) or Gramene (Jaiswal 2011; Youens-Clark et al. 2011) for plants. Curated chemical reaction data, e.g. enzyme names and EC-numbers (~5000 enzymes) linked to publications and protein sequences can be found in BRENDA, which was developed by I. and D. Schomburg at the TU Braunschweig (Scheer et al. 2011). Likewise, the Cazy database provides a classification of carbohydrate active enzymes in sequenced genomes and provides literature references to sequences with experimentally validated function (Cantarel et al. 2009). For transport systems there is TCDB with a specific classification of transport systems and extensive links to the experimental literature (Saier et al. 2009). Specific information on all of the finished and ongoing sequencing projects can be found in

the GOLD dbase (Liolios et al. 2010). Many additional specialized databases exist which will not be mentioned here, but which might be very useful given a particular problem. All of the above-mentioned resources provide the opportunity to search the content using a particular functional description and then allow the retrieval of associated sequences. These sequences can then be compared to the query sequence to evaluate the similarity and a potential overlap in function.

c) Data integration and literature search engines

Investigation of the experimental support is essential to ensure the reliability of a function annotation. This support can be found in two ways, either by the interpretation of new and existing experimental data or by searching the literature. The three most convenient access-points for the experimental literature in the life sciences are Google Scholar (scholar.google.nl), PubMed (Lu 2011) and Web of Science (Thomson; apps.webofknowledge.com). Each of these services has stronger and weaker points, but probably the highest recovery of essential literature is achieved via the former and the latter service. This is related to the fact that part of the literature knowledge-base in the Life Sciences is not represented in PubMed. All three services allow searches forward in time via a 'cited by' option. Again PubMed is the least complete. Other search engines have been developed to search the literature using keywords or sequences directly (examples listed in Table 4) and more efficient next-generation search engines are underway (see (Krallinger, Valencia, and Hirschman 2008)).

d) Public resources of sequences that are considered potentially hazardous.

An important aspect of the risk assessment procedure is the evaluation of the potential hazard that a particular sequence could harbor. In the legislature concerning the use of genetically modified sequences several (protein) functions have been identified as imposing a hazard. These include for instance, antimicrobial resistance, toxicity, virulence-related activity and several other functions. Many associated sequences have been experimentally identified and these have been stored in a manifold of specialized databases (for examples see Tables 5 and 6). One way to assess whether an engineered sequence harbors potential hazard is by comparing that sequence to the entries in these databases. Nevertheless, the interpretation of such an assessment is somewhat problematic given the dispersal of information. For instance, a short survey of the world-wide-web in 2006 retrieved over 30 different web-sites specifically related to data on antimicrobial resistance (Falagas and Karveli 2006). As importantly, it is not very clear how reliable the stored data are due to a lack of clear quality benchmarks and whether the data have been appropriately classified. With regard to the latter, in the case of virulence the actual border between the qualification virulent and non-virulent is clearly a matter of debate (see (Wassenaar and Gaastra 2001))[4].

e) Surfing the web

While searching the world-wide-web one has to be aware that upon uploading a search string, that string becomes public information in the sense that it can be read by others. Although, within the scientific community there is agreement that this should not be done, most commercial parties do not appreciate the risk and perform most analyses on local servers. In fact, many of the public resources allow the extraction of their data to local servers.

4 For example, the two component system QseEF (see Figure 6C) was shown to be important for the virulence of *E. coli EHEC* (Reading et al. 2009). However, this system is present also in the non-virulent *E. coli* strains. The complication arises from the fact that *E. coli EHEC* has genes that confer virulence upon activation by the QseEF two-component system, whereas the non-virulent strains lack these genes.

III) ALGORITHMS USED TO SEARCH AND CLASSIFY SIMILAR SEQUENCES

Sequences with similar molecular function can be detected and collected through sequence and structure comparisons. In the following we will describe the process of searching and comparing sequences and the related bioinformatic tools. Most of the tools that are described can be either accessed directly via the web or can be downloaded to be implemented locally.

a) Pair-wise sequence comparison. The Smith and Waterman algorithm comprises a comprehensive approach to search for similarity between two sequences (Smith and Waterman 1981). Although the algorithm can be speeded up (Farrar 2007), it is often too slow to be used for routine analysis. Instead BLAST (Altschul et al. 1990), later supplanted by the implementations gapped-BLAST and Psi-BLAST (Altschul et al. 1997), is used by the majority of scientists to detect evolutionary and/or functionally related sequences. The algorithm approximates the sensitivity of the Smith and Waterman algorithm but is much faster. Basically, a sequence of interest (query) is compared to all sequences (subjects) present in a reference database in a pair-wise fashion using a scoring matrix such as BLOSUM62. The procedure returns those subject sequences that score below a certain similarity cut-off. The similarity score in terms of E-value relates to the probability that the established relation is found by chance (as explained in (NCBI 2011))[5]. The search can be performed using either a protein or nucleotide sequence as query and a subject database with either protein or nucleotide sequences. Depending on the molecular nature of the sequences that are compared a particular BLAST variant has to be used (e.g. BLASTP for protein against protein and tBLASTN for protein against nucleotide). Besides BLAST there are only a few other popular search algorithms as listed in Table 7.

"Although execution of the search procedure is 'gloriously easy', thanks to many webservers, the outcome requires careful interpretation, like any outcome of a scientific experiment" (Jones and Swindells 2002). The intricacies of BLAST searches have been clearly described by (Altschul and Koonin 1998; Jones and Swindells 2002). One of the main issues relates to the generation and interpretation of the E-value[6] as described below. Standard searches in the NCBI reference sequence database use a treshold E-value of around 10^{-3} - 1. The value has been tested using various sets of sequences with established evolutionary relationship and provides a pretty good recovery of evolutionary related sequences. Nevertheless, the same tests showed that there were often a significant number of false positives and false negatives. Therefore care has to be taken not to interpret the E-value rigidly (Pagni and Jongeneel 2001). The E-value depends on the size of the subject database and the length of the query sequence. In case the database becomes larger, E-values will increase as a result of a higher probability of finding a sequence to be similar just by chance. Vice versa, in case a specialized small database is searched the E-values will be lower, even for sequences that are only remotely similar. In contrast a smaller query sequence will result in a higher E-value even in case of high sequence identity. The E-value also depends critically on the applied scoring matrix. In the case of protein sequence searches, the standard scoring matrix has been created on basis of an average protein composition. Of course, many proteins have a deviating composition. A way to avoid erroneous scoring is by masking such deviating parts of the query sequence during the search. Another way to cope with the problem is to use a scoring matrix that is specifically tailored for the sequences that are being compared (Agrawal and Huang 2011). However, determination of the appropriate scoring matrix for any given query sequence is non-trivial. It is therefore most practical to use a standard scoring matrix and set the E-value cut-off less strict. For instance, when looking for similar protein sequences an initial BLAST search could be performed using a

5 The precise outcome of the calculation will depend on the implementation of the program that is being used.

6 For the sake of convenience the E-value score is often represented in terms of an exponent. 10^{-3} is thus represented as E-3.

A

locus tag	lp_0331	lp_0975	lp_3468	lp_3486	lp_1729	lp_3533	lp_3626	length
lp_0331	0	2.00E-63	2.00E-12	1.00E-13	0.008	1.00E-04	2.00E-04	438
lp_0975	8.00E-67	0	3.00E-12	7.00E-22	0.054	2.00E-05	x	471
lp_3468	4.00E-12	3.00E-12	0	4.00E-80	9.00E-04	3.00E-04	8.00E-07	648
lp_3486	4.00E-12	6.00E-20	1.00E-74	0	0.066	8.00E-05	1.00E-04	652
lp_1729	0.011	0.054	9.00E-04	0.065	0	3.00E-121	4.00E-111	448
lp_3533	2.00E-04	2.00E-05	3.00E-04	8.00E-05	3.00E-121	0	2.00E-135	473
lp_3626	3.00E-04	x	8.00E-07	5.00E-05	4.00E-111	2.00E-135	0	461

B

lp_0331	Begin	End	E-value
Pfam:MFS_1	20	362	1.10e-12
transmembrane	369	391	-

lp_0975	Begin	End	E-value
Pfam:MFS_1	16	384	1.90e-12
Pfam:Sugar_tr	276	387	7.80e-05

lp_3486	Begin	End	E-value
Pfam:MFS_1	17	377	2.30e-08
transmembrane	397	419	-
transmembrane	429	451	-
Pfam:PTS_EIIA_1	492	624	1.40e-47

lp_3468			
Pfam:MFS_1	not identified		
transmembrane	406	428	-
transmembrane	438	460	-
Pfam:PTS_EIIA_	496	627	7.60e-33

lp_1729	Begin	End	E-value
Pfam:MFS_1	22	400	3.50e-25

lp_3533	Begin	End	E-value
Pfam:MFS_1	37	399	2.10e-19

lp_3626	Begin	End	E-value
Pfam:MFS_1	31	399	1.10e-20

C

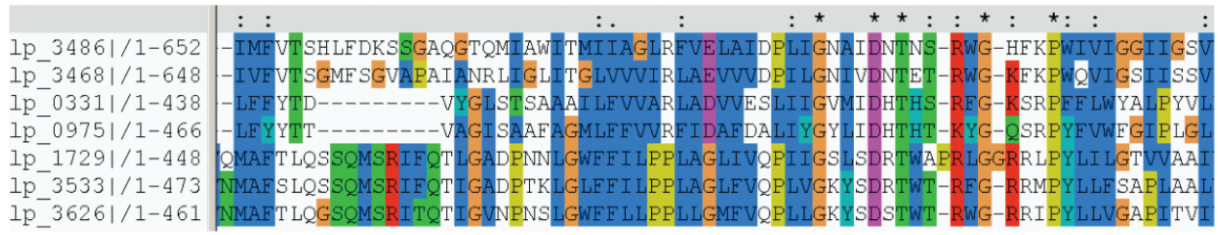


Figure 7. Identification and classification of the GPH-family members of *Lactobacillus plantarum*.
A) The homologs of the GPH-family were collected using an iterative BLAST procedure. The procedure was started with the sequence of MalT (gene *lp_1729*; Figure 2). With an E-value cut-off of E-20, the homologous genes *lp_3533* and *lp_3626* were identified (red cells). With an E-value cut-off of E-5, the gene *lp_3468* was identified, and with that gene also the genes *lp_3486* (blue cells), *lp_0331* and *lp_0965* (green cells). An all-to-all BLAST clearly separated the homologous genes in to three sub-clusters. **B)** The PFAM domains were determined using the profile database of SMART. For six out of seven homologs the PFAM profile was sufficient to recognize the family membership. However, for lp_3468 this was not the case although the BLAST E-value for the comparison lp_3486 vs. lp_3468 clearly implies lp_3468 is a GPH-family member. **C)** Part of the multiple sequence alignment that shows the conservation of the characteristic residues (created using ClustalX).

relatively low E-value cut-off, like 10^{-5} or 10^{-20}. Any subject sequence that fulfills such a criterion can be considered related. In case there are many of these kind of subject sequences the cut off value can be set lower. In case a limited number of sequences is retrieved a second search could be performed with a raised cut-off (e.g. 1 or 10). Even when the overall similarity is low, the query and subject sequences might share clearly conserved short sequence patterns that point at a common molecular function[7].

7The range of E-value cut-offs that is given in this paragraph might seem peculiar considering the magnitude of covered variation. However, in our experience such variation is often needed to retrieve a range of related sequences. This is illustrated for the identification of family members of the gene *malT* in the genome of *Lb. plantarum* in Figure 7.

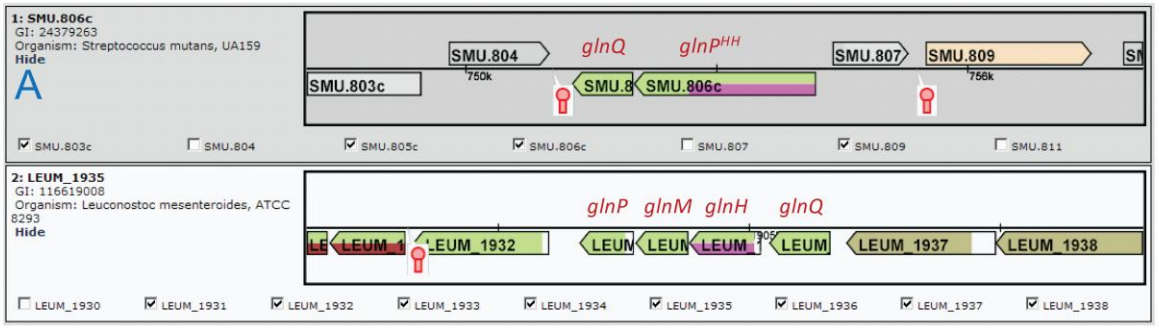
b) Sequence profile comparison.

In general, evolutionary conserved patterns in a sequence are related to the encoded function of that sequence, especially in case the separation in evolutionary time is long (e.g. conservation in genes of distantly related species). These patterns can be described for instance as motifs, profiles, position-specific score matrices or Hidden Markov Models (Schneider et al. 1986; Gribskov, McLachlan, and Eisenberg 1987; Staden 1989b; Staden 1989a; Gribskov 1992; Tatusov, Altschul, and Koonin 1994; Yi and Lander 1994; Bucher et al. 1996; Altschul et al. 1997; Durbin et al. 1999). Basically, all the sequence pattern descriptors start from a multiple alignment of the sequences that are considered equivalent (mostly on basis of the evolutionary relationship). For each position in the multiple sequence alignment, every possible amino acid is assigned a score. If a residue is conserved at a particular position, the corresponding amino acid is assigned a high positive score and others are assigned high negative scores. At weakly conserved positions, all amino acids receive a score near zero. Position-specific scores can also be assigned to potential insertions and deletions (Gribskov 1992; Bucher et al. 1996; Durbin et al. 1999). The various descriptors can be used to score the similarity of a query sequence to a specific pattern.

PSI-BLAST employs the creation of sequence profiles to enhance the recovery of potential functionally related sequences (tutorial in (Bhagwat and Aravind 2007)). The best hits resulting from the initial BLAST similarity search are used to create a sequence profile in the form of a Position Specific Scoring Matrix (PSSM), which is then used to search for similar sequences a second time (Altschul and Koonin 1998; Jones and Swindells 2002). This procedure often works well to identify distantly related sequences, however not necessarily (Altschul and Koonin 1998). The success depends critically on the threshold settings of certain search parameters albeit to a varying extent. Altschul et al. (Altschul and Koonin 1998) provide one example where the increase of the initial similarity threshold was shown essential to improve the recovery of more related sequences whereas they also provide an example where the same increase caused a significant increase in the recovery rate of false positives. Therefore, inspection and curation of the sequence alignment after the initial BLAST search is essential to the success of the PSI-BLAST procedure.

At present, sequence patterns are the main way in which new sequences are functionally classified. Various comprehensive resources of protein sequence patterns or profiles, each related to a particular function, are maintained in the public domain (see Table 8). These include for instance the PFAM (Finn et al. 2010), Interpro (Hunter et al. 2009) and ProDom (Servant et al. 2002; Bru et al. 2005) resources, which are widely being used to annotate functional domains in unknown sequences. As a result of the increasing number of sequenced genomes and experimentally characterized gene products the list of available profiles keeps expanding. Although most of the profile databases can be searched directly using either single or multiple sequences, it is often more convenient to search them via a general server like Interpro (Hunter et al. 2009; Jones et al. 2011) or SMART (Letunic, Doerks, and Bork 2009). In addition, various web-servers like COMA (Margelevicius and Venclovas 2010), COMPASS (Sadreyev et al. 2009), ENA (Leinonen et al. 2011), HHpred (Soding, Biegert, and Lupas 2005; Soding et al. 2006), FASTA, FFAS server (Jaroszewski et al. 2011), or FASTA and SSEARCH (Pearson 1995) offer the opportunity to search similar sequences on basis of sequence patterns and/or structural information from the various profile resources.

Although most profiles directly reflect function, in the sense that the conserved residues are often related to the function, care has to be taken while interpreting the results of a profile-based search (Wong, Maurer-Stroh, and Eisenhaber 2010). The main difficulties relate to overlapping functions (illustrated in Figure 8) and to the appropriateness of the profiles. In case the profiles are based on homologous sequences, it can be that the sequence of several family members has diverged so much from the family consensus that it is not recognized by the profile, whereas it is easily picked up on the basis of a BLAST search with an orthologous or closely related sequence (see Figures 4 and 7). In addition, for a proper evaluation of potential similarity in function it is very important to have



Locus Tag **annotation** **PFAM domains**

B

			SMU.803c	SMU.805c	SMU.806c	SMU.809	LEUM_1931	LEUM_1932	LEUM_1933	LEUM_1934	LEUM_1935	LEUM_1936	LEUM_1937	LEUM_1938
Streptococcus mutans														
SMU.803c	ABC transport?	ABC_tran--ABC_tran	0	9E-13	n	6.9	n	n	n	n	2.3	2E-10	5E-09	1E-10
SMU.805c	ABC transport Gln	ABC_tran	6E-11	7E-146	1.3	2.0	n	n	n	n	n	2E-78	3E-22	2E-23
SMU.806c	ABC transport Gln	SBP_bac_3-(2x)-BPD_transp_1	n	0.42	0	n	0.7	1.6	2E-24	3E-41	3E-15	1.1	1.1	n
SMU.809	ABC subunit exinucl.	ResIII--Helicase_C--UvrB--UVR	9.2	n	n	0	n	n	n	n	n	n	3.2	n
Leuconostoc mesenteroides														
LEUM_1931	ABC transport opp	BPD_transp_1	n	n	0.3	n	2E-177	X	0.03	0.5	4.4	n	n	n
LEUM_1932	ABC transport opp	SBP_bac_5	n	n	1.1	n	n	0	n	n	n	n	n	n
LEUM_1933	ABC transport Gln	BPD_transp_1	n	n	5E-25	n	0.02	n	9E-125	9E-31	n	0.1	n	7.3
LEUM_1934	ABC transport Gln	BPD_transp_1	n	n	7E-42	n	3.4	n	9E-31	7E-122	n	n	n	n
LEUM_1935	ABC transport Gln	SBP_bac_3	1.2	n	9E-16	n	4.1	n	n	n	4E-171	n	n	n
LEUM_1936	ABC transport Gln	ABC_tran	5E-09	2E-78	0.3	3.1	n	n	0.1	n	n	9E-146	8E-23	2E-18
LEUM_1937	ABC transport mdrug	ABC_membrane--ABC_tran	2E-07	7E-22	0.04	2.8	n	n	n	0.1	n	2E-22	0	9E-63
LEUM_1938	ABC transport mdrug	ABC_membrane--ABC_tran	4E-08	4E-23	n	n	n	n	n	n	n	6E-18	9E-63	0

Figure 8. BLAST and profile comparison of the ABC transport system for glutamine between the Lactobacilli: *Streptococcus mutans* and *Leuconostoc mesenteroides*.

A) The gene context for the genes encoding the transport system in both species is given. The system consists of three components: a characteristic ATP-binding protein/domain (PFAM: ABC_tran), a permease protein/domain (PFAM: BPD_transp_1 in the case of polar amino acids), and a substrate binding protein/domain (PFAM: SBP_bac_3 in the case of polar amino acids). In *Streptococcus mutans* these components are encoded by two genes *glnQ* (ATP binding) and *glnP*^{HH} (permease and two substrate binding domains), whereas in *Leuconostoc mesenteroides* they are encoded in four separate genes *glnQ* (ATP binding), *glnH* (substrate binding), *glnM* (permease) and *glnP* (permease). The surrounding genes also encode mostly components of different ABC transport systems.

B) One-to-one BLAST E-values and PFAM domains associated with the ABC transport encoding genes. It is clear from the E-values that the ATP-binding domain of various systems has common features (light blue boxes). Yet the similarity between closely related domains is also apparent from the E-values. The simililarity in terms of E-values between the permeases is only clear for the same substrate (purple boxes). Yet, on basis of the PFAM sequence profiles it can be deduced that the oligopeptide permease (opp) and the glutamine permease (Gln) are homologs. In the case of the glutamine substrate binding protein/domain the one-to-one E-values are considerably higher (red-brown boxes) than for the glutamine permease or the related ATP-binding component, although they all represent orthologs. The large variety in one-to-one E-values for orthologous sequences simply relates to the variety in the number of residues that needs to be conserved to retain functionality.

an idea of which sequence elements are present, like transmembrane helices or leader peptides. Various tools have been constructed to identify the latter elements in a query sequence (listed in Table 8). They include SignalP for leader-peptides (Bendtsen et al. 2004; Petersen et al. 2011), and TMHMM (Krogh et al. 2001) for transmembrane helices. Moreover, several specific databases exist with transmembrane profiles like (Arai et al. 2004) or transmembrane proteins like TCDB (Saier, Tran, and Barabote 2006). In all profile tools the given parameter set has been optimized to provide the best search results in the hands of non-experts.

c) Modeling protein structure

For many proteins and protein-families structural data can be exploited to evaluate the potential functional relevance of differences in sequence. In the case of a small change in a protein of known structure, like a change in a single amino acid residue, the prediction is mostly straightforward (see (Venselaar et al. 2010b; Worth, Preissner, and Blundell 2011)). The process of extrapolating a known protein structure to a putative structure for a similar protein, which is called homology modeling, is also well do-able especially for sequences with high identity (see (Krieger, Nabuurs, and Vriend 2003; Guo, Ellrott, and Xu 2008; Venselaar et al. 2010a)). The generation of reliable *ab initio* protein structure predictions has proved harder. Nevertheless, the quality of the predictions has increased considerably over the years (Fischer 2006; Dill et al. 2007). The quality of homology-based and ab initio methods are tested by the scientific community in the two-yearly CASP competition (Kryshtafovych et al. 2009; Moult et al. 2011). Currently, many of the methods are available as a web-service (see Table 8). The practicalities of protein structure prediction are discussed in (Watson, Laskowski, and Thornton 2005; Mazumder and Vasudevan 2008; Venselaar et al. 2010a; Pavlopoulou and Michalopoulos 2011).

Compared to protein structure prediction, RNA structure prediction is still in its infancy. This is signified clearly by lesser tools and servers and also by a far more confined body of literature. However, the changed perception of the metabolic role of RNA, namely as another major player in the control of cellular activity, will probably spur the progress in this field of expertise. A nice overview of the current state of knowledge is presented in (Masquida, Beckert, and Jossinet 2010; Westhof, Masquida, and Jossinet 2011).

d) Multiple sequence alignment

The functional similarity of sequences is generally derived from an analysis of their kinship. And the familiar relationship between sequences is usually evaluated on basis of an alignment of a set of multiple related sequences (Levasseur et al. 2008; Wong, Suchard, and Huelsenbeck 2008; Thompson et al. 2011). Whereas in a pair-wise alignment simple and tractable algorithms can be used to provide the best result, direct extension towards the alignment of multiple sequences is non-trivial. Most popular approaches employ a so-called progressive algorithm (Feng and Doolittle 1987) which has been extended by a refinement strategy, like using iterations or profiles. These approaches include for instance CLUSTALX/W (Larkin et al. 2007), Kalign (Lassmann, Frings, and Sonnhammer 2009), MAFFT (Katoh and Toh 2008), MUSCLE (Edgar 2004a), ProbCons (Do et al. 2005) and T-Coffee (Notredame, Higgins, and Heringa 2000; Di Tommaso et al. 2011). Unfortunately, none of them is best-suited for all alignment problems and the preferred method depends upon the characteristics of the particular set of sequences that is to be aligned in terms of the number of sequences, their length, their evolutionary distance, the variability in composition (e.g. varying number of domains) and so on. For protein alignments several benchmark sets of sequences have been created and have been used to determine which method works best in general (Blackshields et al. 2006; Aniba, Poch, and Thompson 2010; Thompson et al. 2011). Performance tests (Pei 2008; Thompson et al. 2011) show that on average all popular aligners provide reasonable alignments for sequences that are relatively similar (>30% similarity) but become worse for more divergent sequences or for larger sets. Recently, Clustal has been upgraded so that it can handle very large sets of sequences. The new tool is called Clustal-Omega (Sievers et al. 2011).

A listing of popular multiple sequence alignment tools is provided in Table 9. All of the tools can be downloaded and locally installed, whereas some are also available via a web-server (e.g. via sequence analysis tool-boxes such as Jalview (Waterhouse et al. 2009) and MEGA (Tamura et al. 2011)). Two important features that contribute to the practical use of the various tools are the ease with which they can be installed and used and the time they need to generate an alignment. Because of that, for routine use the fast methods are preferred. For instance, CLUSTALW (Larkin et al. 2007) is the most used aligner probably because it is easy to download, install and use, even though it is not the most accurate program (Aniba, Poch, and Thompson 2010). Also MUSCLE (Edgar 2004) is easy to download and is fast. However, it runs on command line (DOS or UNIX). Kalign (Lassmann, Frings, and Sonnhammer 2009) is the fastest, performs well, but is less easy to install. The most accurate fast tool is MAFFT (Katoh and Toh 2008), which is easy to download and install, and runs on command line (DOS or UNIX). Protein (and nucleotide) alignments can be improved in case structural information is available. Tools like Expresso (Di Tommaso et al. 2011), PRALINE (Brandt and Heringa 2011), PROMALS(3D) (Pei and Grishin 2007; Pei, Kim, and Grishin 2008) or SPEM(3D) (Zhou and Zhou 2005) can include this kind of information to guide the result.

e) Phylogeny and classification

As discussed earlier, the evolutionary relationship between sequences can be taken as indicative of their functional similarity (Eisen and Wu 2002). The study of the evolutionary history of genetic material is referred to as phylogenetics. The historical relationship between genes and proteins can be represented in a phylogenetic tree (Whelan 2008). The construction of historically reliable trees for gene and protein sequences is a research field of its own (see books by (Felsenstein 2003; Lemey, Salemi, and Vandamme 2009; Hall 2011; Huson, Rupp, and Scornavacca 2011). Especially, defining the root of the tree can be highly problematic. Fortunately, appropriate rooting is often not needed to establish whether sequences share a close evolutionary relationship (Morrison 1996).

Most detailed phylogenetic analyses are based on the multiple sequence alignment. There are several basic approaches to construct a phylogenetic tree out of the alignment (see (Morrison 1996; Holder and Lewis 2003)). Constructive methods involve hierarchical clustering based on the distance matrix that was generated during the alignment (i.e. from the pairwise comparison of sequences), whereas search methods involve the reconstruction of a putative ancestor state(s) on basis of the observed per position variation in the alignment and a statistical or numerical analysis to select the optimal tree. Examples of the first approach are the popular Neighbor Joining algorithm and UPGMA, whereas examples of the second are maximum likelihood and maximum parsimony (Morrison 1996; Gascuel and Steel 2006; Whelan 2008). In addition, Bayesian analyses have been introduced for the purpose of tree generation (see (Holder and Lewis 2003)). Although each of the methods has particular advantages and disadvantages the Neighbor Joining algorithm is most commonly used for fast routine analyses. In general Neighbor Joining suffices to cluster orthologous sequences correctly, where the correctness for the observed clustering becomes apparent from the associated statistical support, like acquired by using bootstrapping (technique explained in (Varian 2005); Figure 9). Of course also more complicated methods have been developed to analyze the reliability and comparability of sets of trees (e.g. (Hillis, Heath, and St John 2005; Arnaoudova et al. 2010)).

For the analysis of individual cases, a clear visualization of the trees and of the genomic context is essential to enable rapid interpretation. Many tools and servers have been developed to achieve clear visualizations (examples given in Table10 and 11). Easy to use programs/servers for tree-viewing are Dendroscope (Huson et al. 2007), LOFT (van der Heijden et al. 2007) and iTOL (Letunic and Bork 2007). Often used servers for the analysis of gene context include: Artemis (Carver et al. 2008), Microbial Genome Viewer (Kerkhoven et al. 2004), Microscope (Vallenet et al. 2009), phiGENOME (phages (Stano and Klucar 2011)), UCSC Archaeal Genome Browser (Chan et al. 2011) and NCBI Map viewer (Wolfsberg 2011).

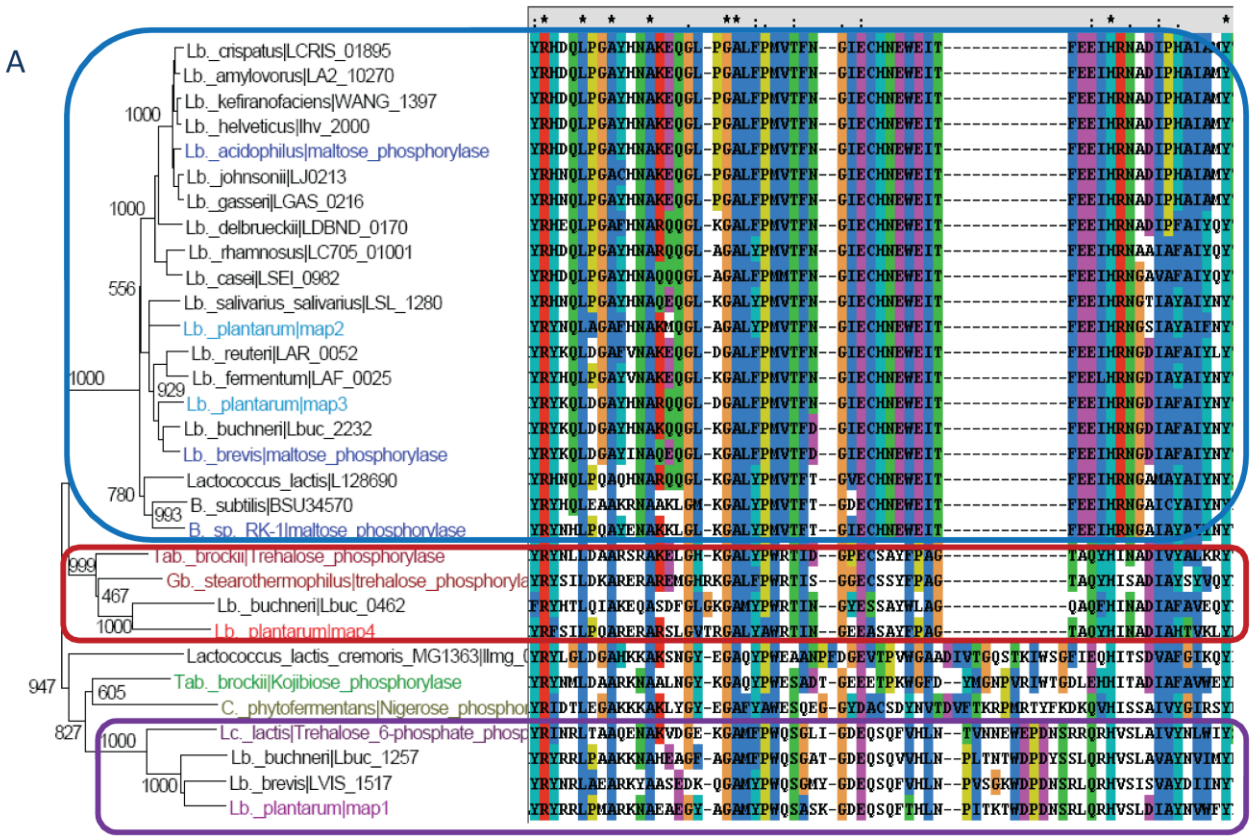


Figure 9. Function annotation of the Glycosyl Hydrolase family 65 homologs of *Lb. plantarum*.
The genes encoding putative maltose phosphorylases in *L. plantarum* (*map2* and *map3* in Figure 2) were used to search for homologs in all Lactobacillus genomes (the process is described in detail in Appendix A). The protein sequences were collected and sequences with experimentally validated function, taken from the CAZY database, were added (indicated in color). All sequences were aligned and a Neighbor Joining tree was generated on basis of the alignment. **A)** Neighbor Joining tree and multiple sequence alignment for the Glycosyl Hydrolase family 65 homologs of *L. plantarum*. **B)** Putative annotation for the genes map1-4 of *L. plantarum* as derived from the clusters in the Neighbor Joining tree.
The clusters show a very high bootstrap support (>999/1000) and the sequences within the clusters are well-aligned (no gaps within the clusters). These two observations are strong indicators that the sequences within the clusters are orthologous and thus share the same molecular function. In the case of *map2* and *map3*, they are both orthologous to the gene encoding maltose phosphorylase in *B. subtilis* (Inoue et al. 2002). In addition, there is experimental evidence for maltose phosphorylase activity for several other sequences within the cluster (Huwel et al. 1997; Nakai et al. 2009). Map 4 is orthologous to trehalose phosphorylase of *Thermoanaerobacter brockii,* whereas Map1 seems orthologous to trehalose 6-P phosphorylase of *L. lactis IL1403.* The part of the alignment that is depicted represents the region between loop 3 and 4 (structure in (Egloff et al. 2001), which was shown to affect substrate specificity (Nakai et al. 2010).

For the analysis (i.e. classification) of many sequences at the same time various high throughput methods have been developed (examples given in Table 12 from (Kuzniar et al. 2008)). These methods can be either 'graph-based' like in the case of COG (Tatusov et al. 2003) and eggNOG (Powell et al. 2012), be 'tree-based' like in the case of LOFT (van der Heijden et al. 2007), or hybrids of the two. By necessity the results of these methods will be less accurate than the results of the more laborious procedure described earlier. However, in many cases also the high throughput methods suffice to cluster evolutionary closely related sequences reliably

(Koonin 2005; Trachana et al. 2011). In fact, the problematic cases are mostly obvious. They relate to sequences with high identity to many others (i.e. larger gene/protein families with multiple homologs per genome), and on the other hand, to sequences that share low identity with only few sequences. The former group often is related to transport, carbohydrate metabolism and transcription regulation, whereas the latter group includes many of the so-called 'hypotheticals'.

IV) ANNOTATION OF FUNCTION

In case there is an orthologous relationship between sequences and the genomic context is conserved we may in general assume functional equivalency (i.e. same molecular function, same biological role). In this section we will describe the process of the attribution and evaluation of a putative annotation for a particular sequence by application of the methods, tools and services described in the previous sections. It is important to note that a high sequence identity (or membership of the same orthologous group), although it is most often the case, does not necessarily mean functional equivalency (Rost 2002). At the same time it is important not to take the annotations present in the public domain for granted. Soon after the publication of the first complete genome sequences it was already noted that the publicly available data contain many annotation errors (Brenner 1999; Devos and Valencia 2000; Iliopoulos et al. 2001). As a consequence, the results of an automated annotation procedure have to be, what is called 'curated' (Francke, Siezen, and Teusink 2005). In the process of curation, the annotation has to be somehow linked to experimental evidence as present in literature. However, also that what is in text presented as 'experimentally verified' should be checked with the actual data to establish whether the claims are justified, as is nicely illustrated in the paper by (Iyer et al. 2001).

a) Inference of molecular function: transfer of annotation

A basic manual annotation strategy for sequences of unknown function (query) consists of a number of semi-independent steps as illustrated in Figure 9 and Appendix A (Bork et al. 1998; Whisstock and Lesk 2003). The first steps involve the recovery of homologous sequences (subjects) via a BLAST search, the multiple alignment of those sequences, followed by the construction of a tree to establish the putative evolutionary relationship between them. The following steps involve a comparison of the genomic context of all recovered sequences and the collection of sequence associated information using either sequence profiles or group classifications (like COGs) and literature.

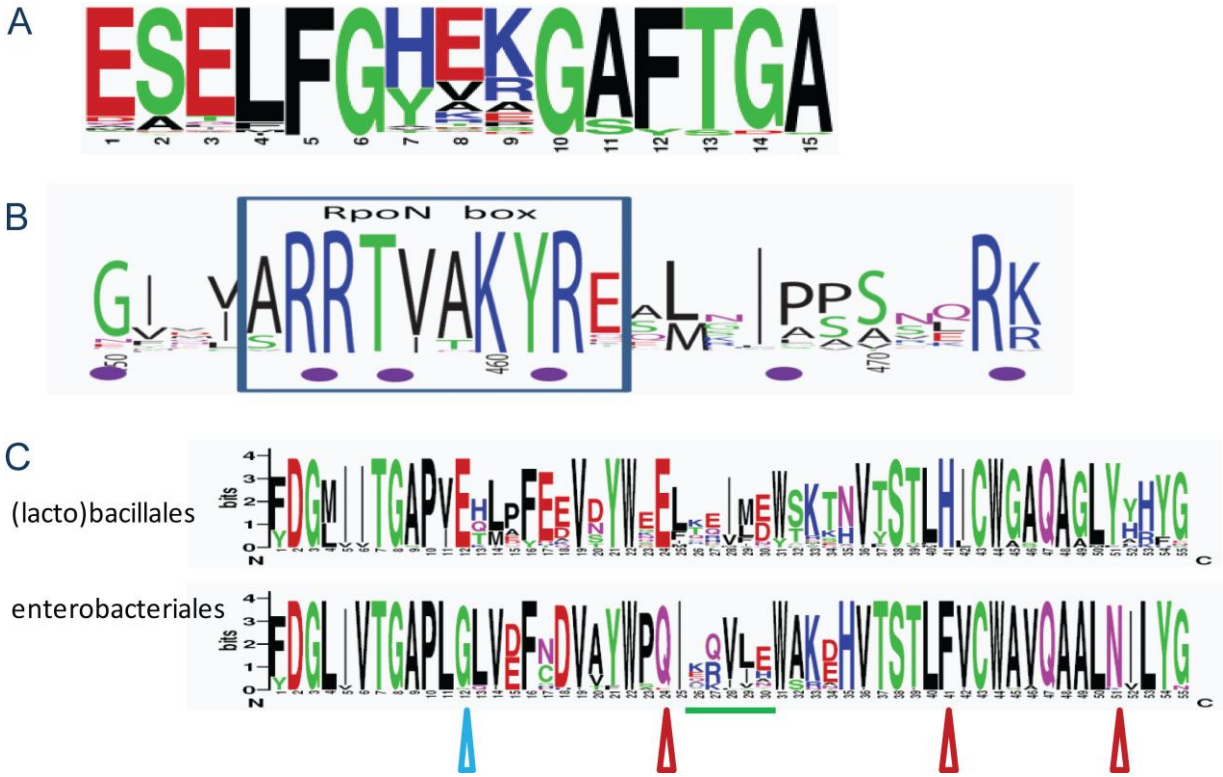


Figure 10. The natural variability of protein sequence and its reflection of function.
In general one may assume for a set of orthologous/homologous sequences that those residues that have been conserved contribute more to the molecular function than those that have not been conserved. Vice versa one may assume that a change in the former will have an effect on function, whereas a change in the latter will not. Moreover a distinction can be made between those residues that are conserved between all homologs, i.e. which will mostly confer general functional characteristics, and residues that are conserved between a subset of homologs, i.e. which will confer more specific characteristics.
A) Conservation of the characteristic sigma-54 interaction sequence within 5000 sigma-54 related Enhancer Binding Proteins (EBP). The part of the EBP sequence that is responsible for the direct interaction with sigma-54 is highly conserved and a single amino acid change can abolish the interaction (see Figure 4). The level of conservation is rare given the fact that sequences are derived from species representing very different domains of the kingdom bacteria. One may assume that especially changes in the fully conserved residues (nr 5, 6 10) will affect the binding of the EBP to sigma-54. Nevertheless, not all changes in these highly conserved residues appear to abolish the activity (see refs in (Francke et al. 2011)). It has been shown that replacement of the Phe at position 12 (Zhang et al. 2009) and the Thr at position 13 (Dago et al. 2007) have a huge negative effect on activity. **B)** Conservation of the RpoN-box in the sigma-54 sequence throughout the bacterial kingdom. The RpoN box is responsible for the interaction of the sigma factor with the promoter DNA. In this case the level of sequence conservation corresponds nicely with the importance for the maintenance of function as was shown in an Ala-Cys scanning mutagenesis study (Xiao et al. 2009). Mutagenesis of the residue-pairs marked by the purple dots abolished the DNA-binding (almost) completely. Exchange of the other residues hardly affected function. This is in line with the observed lack of conservation. **C)** Comparison of the *metA* gene, encoding homoserine O-succinyl/acetyl transferase, between Enterobacteriaceae and Bacillaceae. The stretch of sequence containing the residues that compose the acetyl/succinyl binding site is shown. The enterobacterial enzyme is a homoserine O-succinyl transferase (Born and Blanchard 1999), whereas a single amino acid mutation (at position 111 in the reference structure as indicated by the blue arrow) makes the *Bacillus cereus* ortholog a homoserine O-acetyltransferase (Zubieta et al. 2008). The residues indicated by arrows display a conserved difference between the metA sequence of enterobacteriales and (Lacto)Bacillales. Therefore these residues represent prime candidates to confer potential differences in function, whereas for instance the residues underlined in green show such a high degree of variability within the subsets that a change in any of them probably has no effect on the functional properties. MetA provides a perfect example of the divergent evolution of some orthologs.

In principle, the annotation can be transferred directly from a subject to the query in case the observations made in the three initial steps point in to the same direction, namely of clear orthology, a conserved genomic context and identical sequence profiles. In case, the correspondence between the data related to the query and subject sequences is reduced, the function is less predictable. For instance, one can infer that the function and role of non-orthologous homologs that share a conserved genomic context will be highly similar, and that, in case the context is not conserved, the function will be probably similar. The reliability of the information transfer thus increases with an increasing number of independent observations in support.

Important clues concerning protein function can come from structural data (Whisstock and Lesk 2003; Mazumder and Vasudevan 2008; Fischer 2006; Watson, Laskowski, and Thornton 2005). Many structures of proteins with unknown function have been solved on an individual basis, but more and more, also in a high throughput manner (Yakunin et al. 2004). Structural alignment with proteins of known function then can help to infer putative functional similarity (Mayr, Domingues, and Lackner 2007; Liu, Srivastava, and Zhang 2011). As we have mentioned before, in the case structural data are available homology modeling can be used to pinpoint functionally important residues (Venselaar et al. 2010a; Venselaar et al. 2010b). Similarly, a comparison of the conservation patterns displayed by conserved residues between groups of paralogous sequences can be used to pinpoint those residues that confer functional specificity (Bharatham, Zhang, and Mihalek 2011). In fact one-residue changes that alter the protein specificity can be traced that way (see (Kumar, Henikoff, and Ng 2009) and Figure 10).

The inference procedure described above can be applied to sequences that evolve divergently (majority). However, there are some well-studied cases of convergence, where sequences that share no common history have evolved to perform the same task, i.e. to have the same molecular function and biological role. In the absence of experimental data there is no way in which a single sequence-based analysis can detect such an analogy between sequences. Nevertheless, genes present in the gene context might point in the right direction as well as the regulatory elements that are associated to the particular genes. A perfect example can be found for the enzyme that catalyzes the hydrolysis of lactose in to its constituent sugars glucose and galactose, beta-galactosidase as depicted in Figure 11.

Correlations found in high throughput experimental data, like transcriptomic, proteomic and/or metabolomic data, can be used to support the function annotation of a particular sequence (e.g. (Clare and King 2002)). Nevertheless and paradoxically, although this kind of experimental data represent the actual state of cells/organisms, correlations between the presence of certain sequences and the observed behavior can not be taken blindly as indicative of the role of those sequences because many of the observed effects may be indirect. In fact, a thorough sequence analysis and/or dedicated experimental investigation of the molecular mechanism will often be needed to prove a causal relationship between the observed phenomena and the (in-)activity of certain gene products. A tool that approaches such kind of integrative analysis for a large number of sequences present in the public domain is STRING (Szklarczyk et al. 2011).

b) High throughput annotation

It is clear that a purely manual strategy is far too laborious to achieve annotation for large sets of sequences, like for instance in the case of a newly sequenced genome. Therefore many function recognition and annotation pipelines have been developed (see (Siezen and van Hijum 2011) and Table 13). Basically, these pipelines represent an automated version of the process described in the previous paragraphs. Due to the automation the reliability of the annotation will be less clear. Nevertheless, much of the function annotation found in the public domain is derived from using such pipelines. The reliability can be enhanced by relating the newly sequenced genome to the content of a curated database (e.g. like the EnzymeDetector pipeline does (Quester and Schomburg 2011)).

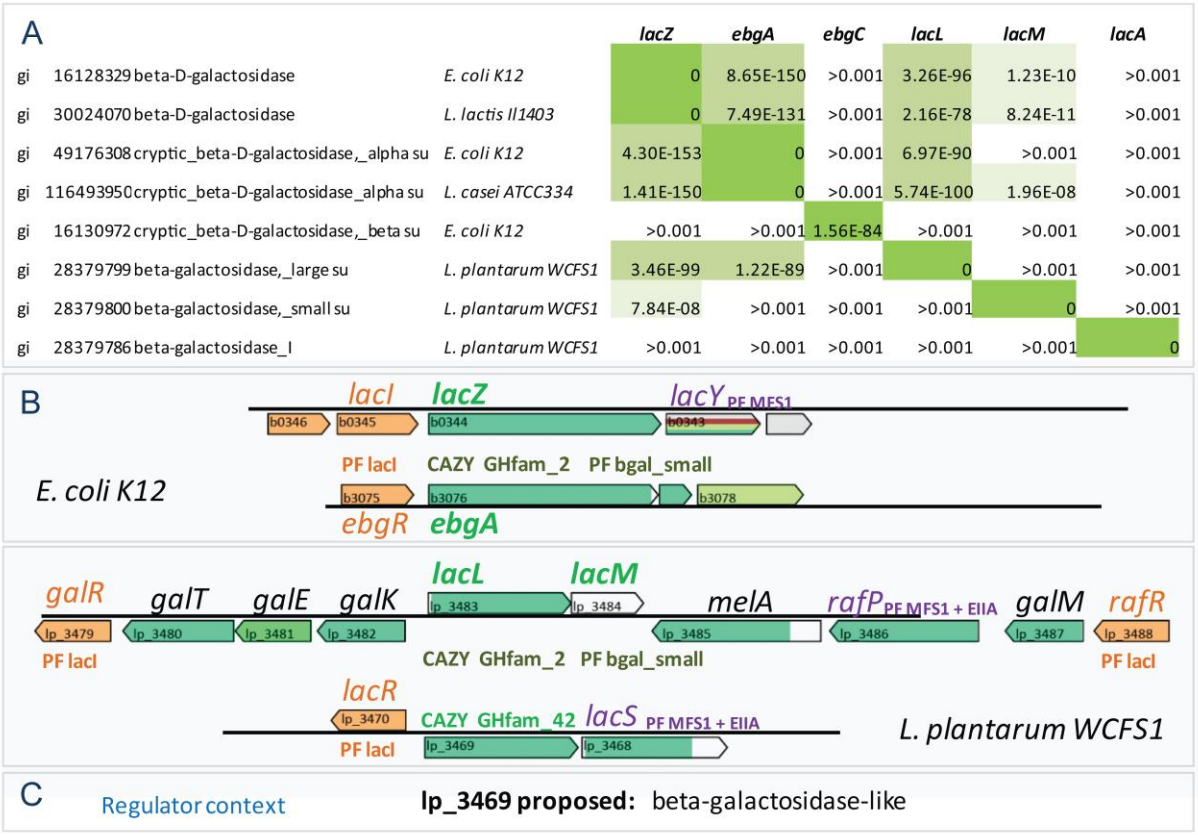


Figure 11. Function annotation of the beta-galactosidases in *Escherichia coli* and *Lactobacillus plantarum*.

A) Comparison of the sequence similarity of the beta-galactosidases of *E. coli* and *L. plantarum* using BLAST. **B)** Gene context of the corresponding genes. The associated regulators (in orange) are all members of the LacI-family. In fact, *E. coli* EbgR and *L. plantarum* LacR, RafR and GalR belong to a separate clade within the family (Francke et al. 2008). Their relatedness suggests they will respond to very similar signal molecules. **C)** Function annotation of *lp_3469* on basis of the genomic context. In the absence of additional experimental knowledge one could provide the gene *lp_3469* with a putative function annotation on basis of the two genomically associated genes. The first gene, *lacS*, encodes a lactose proton symporter (purple; see Figure 7) and the second a regulator of the lacI-family with an inducer related to lactose (see B). In addition, the encoded protein sequence is characteristic for a glycoside hydrolase. Therefore, *lp_3469* encodes a putitative lactose hydrolase, i.e. a beta-galactosidase. Indeed, the orthologs found in *L. acidophilus* and L. reuterii (gene *lacA*) were shown to bind similar substrates as LacZ and LacLM, although with different specificity (Schwab, Sorensen, and Ganzle 2010; Andersen et al. 2011).

The production of galacto-oligosaccharides using microbial beta-galactosidases is currently well-studied in the field of functional foods (Park and Oh 2010). In *Escherichia coli* a gene encoding beta-galactosidase: *lacZ,* was described first by Joshua Lederberg in 1948 (Lederberg 1948). And the regulation of the gene has become the archetype of the organization of prokaryotic transcription regulation. It took 25 years before a second beta-galactosidase encoding gene was discovered in *E.coli* (Campbell, Lengyel, and Langridge 1973). This gene was designated *ebgA,* from evolved beta-galactosidase. The discovery resulted in the classic study (designation by (Dean 2010)) of molecular evolution (review in (Hall 2003)). The BLAST E-values comply with the assertion that both genes have evolved from a common ancestor. In many lactobacilli a third closely-related variant is found, *lacLM.* In some Lactobacilli (*L. delbrueckii* and *L. salivarius*) the protein is encoded by a single gene. However, in most Lactobacilli the protein is encoded by two neighboring genes (probably the result of gene fission) and the active protein is a heterodimer (Nguyen et al. 2007a). It is the LacLM protein that is exploited in the biotechnological applications (Nguyen et al. 2007b; Liu et al. 2011). Like *E. coli*, various Lactobacilli have a second beta-galactosidase encoding gene, *lacA*. However, this gene has a completely different evolutionary origin and thus represents a functional analog.

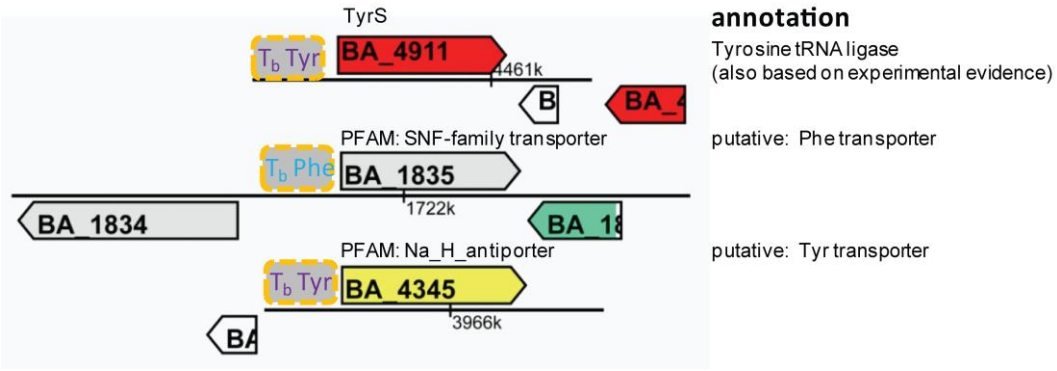


Figure 12. Function annotation of amino acid transporters in *Bacillus anthracis* on basis of the amino acid specificity of the associated regulatory elements.
In many Bacilli and Lactobacilli a number of genes related to amino acid metabolism and transport is preceded by a regulatory sequence called a T-box riboswitch (Gutierrez-Preciado et al. 2009). In amino acid-rich conditions the T-box forms a terminator structure, thereby preventing transcription of the genes downstream. Transcription can ensue after the formation of the anti-terminator structure which is induced upon binding of uncharged t-RNA to the so-called specifier codon. As the name implies the specifier codon recognizes only one kind of t-RNA and thus makes the response amino-acid specific. The amino-acid specificity of a T-box thus can thus be taken as an indicator of the molecular function of the gene downstream (Wels et al. 2008).

c) Enzyme promiscuity

The function annotation of genes and proteins is often perceived as describing a singular trait. As we have pointed out earlier, this singular view should be corrected because most of these molecules have multiple roles. Moreover, with regards to function annotation of proteins there is an additional layer of complexity which is related to the fact that several substrates look and behave similarly. Indeed, for many proteins it has been observed that they can act on multiple substrates. This phenomenon is nowadays referred to as 'enzyme promiscuity' (for a comprehensive review see (Khersonsky and Tawfik 2010)). Also in the case of transport most individual systems have been associated to multiple substrates (see for the bacterial PTS e.g. (Postma, Lengeler, and Jacobson 1993)). Fortunately for many enzymes information about the additional substrates can be easily found, for instance in the databases of KEGG (Kanehisa et al. 2008) and BRENDA (Scheer et al. 2011). Enzyme promiscuity has gained interest because of its potential application in biotechnology (Hult and Berglund 2007). Efforts to trace and predict new promiscuous functions focus at both the protein sequence level (Carbonell and Faulon 2010) and at the level of the substrates (Nath and Atkins 2008; Kim, Bolton, and Bryant 2011), the latter often by applying a simple similarity measure called the 'Tanimoto similarity' (Rogers and Tanimoto 1960). In addition, compound similarity measures can also be used to infer properties like toxicity, as exemplified recently in *E. coli* (Planson et al. 2011) and *Plasmodium* (Gonzalez-Diaz et al. 2011).

d) Inference of biological role: reconstruction and modeling

The cellular system as a whole, as represented in its biochemical pathways, signal transduction routes, the structure of cellular molecular machines, and the composition of the cellular structures, provides the functional context of a gene/protein sequence. Thus a reconstruction of the system will add to the potential identification of the biological role of the gene/protein sequence. The conversion of the reconstruction into a model will ultimately enable the quantification of that role (Reed et al. 2006). The process of sequence-based reconstruction in principle involves exactly the same methods and procedures as described earlier, the only difference being that now the acquired information is put into a broader functional context (Francke, Siezen, and Teusink 2005). Various pipelines exist that allow reconstruction of the cellular metabolism on basis of its genome (like e.g. IMG (Markowitz et al. 2012), RAST (Aziz et al. 2008) and PathwayTools (Karp et al. 2010)). A large number of reviews have appeared describing the process and pitfalls of metabolic reconstruction (e.g. (Francke, Siezen, and Teusink 2005; Duarte et al. 2007; Ma et al. 2007; Satish

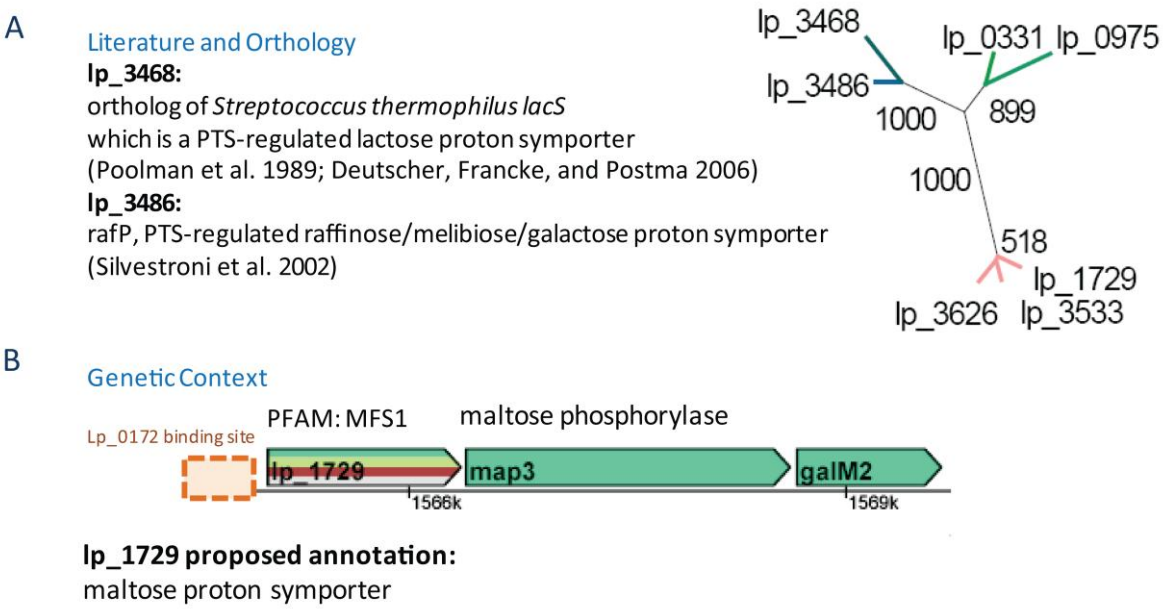


Figure 13. Function annotation of the GPH-family transporters of *Lactobacillus plantarum*.
The genes encoding the GPH-family transporters were collected and classified as illustrated in Figure 7. A bootstrapped neighbor-joining tree was generated on basis of the multiple sequence alignment. The tree shows three separate clusters. The cluster consisting of *lp_3468* and *lp_3486* is special in the sense that the encoded transport proteins contain a PTS-related signal domain, the status of which affects the mode of action of the transport protein. **A)** Function annotation of genes *lp_3468* and *lp_3486*. The annotation of *lp_3486* can be based on primary experimental evidence, whereas the annotation of *lp_3468* can be inferred on basis of orthology with *lacS* of *S. thermophilus* and the experimental evidence presented for the molecular function of that gene/protein. **B)** Function annotation of the gene *lp_1729*. The annotation as a maltose proton symporter is based on the genomic association with maltose phosphorylase (*map3*, Figure 9) and the general functional trait of GPH-family homologs, which is to transport (di-)saccharides via proton symport (Poolman et al. 1996). Moreover, the operon is preceded by a binding-site for the maltose-activated TF Lp_0172 (Francke et al. 2008).

Kumar, Dasika, and Maranas 2007; Osterlund, Nookaew, and Nielsen 2011; Pitkanen, Rousu, and Ukkonen 2010)) and the creation and application of metabolic models thereof (e.g. (Price et al. 2003; Durot, Bourguignon, and Schachter 2009; Oberhardt, Palsson, and Papin 2009; Blazeck and Alper 2010; Chen et al. 2011)).

Reconstruction of the regulatory network that is associated to the metabolic enzymes could add to the understanding of the dynamics of an organism's physiology and the role of particular signaling metabolites therein. These reconstructions can be based both on genome sequence data (Rodionov 2007) and on high throughput data, but are currently mainly based on the latter (Herrgard, Covert, and Palsson 2004). Unfortunately, every network inference that is based on high throughput data alone suffers from the fact that the system is experimentally underdetermined (De Smet and Marchal 2010). Latter problem can be tackled by directing the reconstruction on basis of additional considerations. One should realize that the direction of the ultimate outcome then will be predetermined. The additional considerations include for instance binding-site definitions based on purely *in silico* methods but also supervised footprinting approaches (Stormo 2000; Bulyk 2003; Thompson, Rouchka, and Lawrence 2003; Alkema, Lenhard, and Wasserman 2004; Van Hellemont et al. 2005; Wels et al. 2006; Okumura et al. 2007; Francke et al. 2008; Nain, Sahi, and Kumar 2011). Two examples of enhanced annotation based on knowledge of the regulatory connections are given in Figures 12 and 13. Despite the above, the (automated) reconstruction methods are still at a stage of infancy (Fuellen 2011).

3) CONSIDERATIONS REGARDING THE POSSIBLE APPLICATION OF BIOINFORMATICS IN GMO RISK-ASSESSMENT STRATEGIES

There is considerable variability in the extent to which genetic material is engineered in different biotechnological applications (Carr and Church 2009). And thus there is a considerable variability in the ease and reliability with which the potential effects can be predicted. In the simplest case, to improve the yields in the industrial production of a particular protein one might adapt the codon usage within the related gene (thus without affecting the protein sequence). Then the most important effect that has to be taken in to account in the prediction is the increased local concentration of the particular protein. A far more complicated case is presented in (Fisher et al. 2011), where a library of random 102 amino acid-long sequences were tested *in vivo* for their ability to rescue particular knockout phenotypes in *E. coli*. To formulate the expected outcome and/or risk of such an experiment, the random sequences have to be attributed a putative molecular function and the potential effects of the presence of these functions inside the cell has to be evaluated. In this Chapter we will illustrate for a few examples how the bioinformatics-based analyses described in the former chapter can be used to answer some of the questions related to these kind of applications. The description will start with the simple case of introducing a limited number of changes in a sequence of known molecular function and will end with more complicated cases.

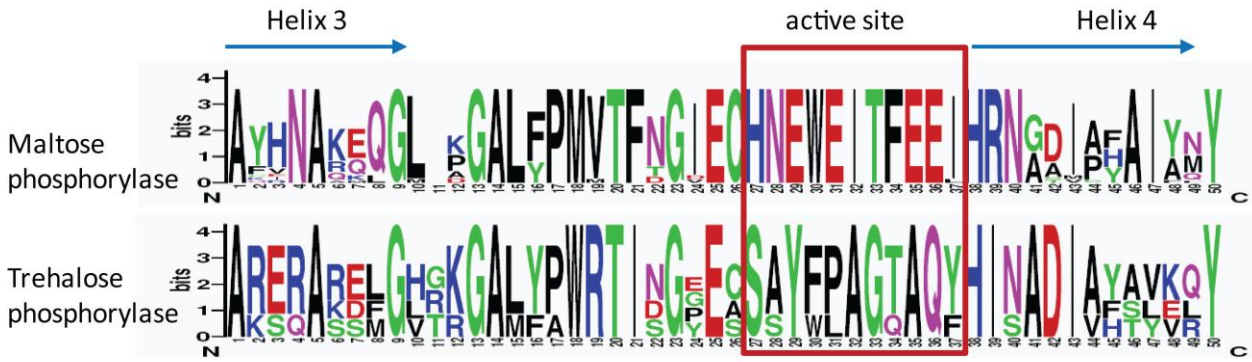


Figure 14. Rational engineering of *Lactobacillus acidophilus* maltose phosphorylase.
Recently the maltose phosphorylase of *L. acidophilus* was successfully engineered into a trehalose or kojibiose dual specifity phosphorylase using only a few amino acid residue replacements that were implied by homology modeling of the protein (Nakai et al. 2010). First the protein sequence of the maltose phosphorylase of *L. acidophilus* was aligned with that of *L. brevis* and then, from the alignment, the sequence was superimposed on the crystal structure of the *L. brevis* protein (Egloff et al. 2001), together with the substrate acarbose taken from the crystal structure of substrate-bound glucoamylase (Aleshin et al. 2003). The model implied the region between loop 3 and loop 4 to be important for substrate specificity (Nakai et al. 2010). The Figure depicts this region for the average sequence of the maltose phosphorylase orthologs and the trehalose phosphorylase orthologs as derived in Figure 9. The amino acid residues in loop 3 and loop 4 show conservation for the two orthologous groups (related to family traits), whereas the residues assigned to the active site show a conservation pattern that is clearly different between the groups (red-brown box; related to specific function). Mutation of the related residues in the maltose phosphorylase of *L. acidophilus* imposed trehalose activity on the enzyme. The resulting activity was dual activity (both on maltose and trehalose), whereas the original maltose phosphorylase had a strict substrate specificity (Nakai et al. 2009). Moreover, the specific activity was much lower than in the original enzyme. The latter phenomenon is often observed in the rational re-design of enzyme activity (Gerlt and Babbitt 2009) and is probably related to the fact that many residues contribute to the specificity of the interactions albeit in a more indirect way (e.g. via subtle effects on the overall structure).

I) VARIATION IN PROTEIN SEQUENCE

a) The potential effect of amino acid substitutions and the natural variability argument

Many studies start with a sequence of known function and introduce a few changes to subtly change the molecular function. The evaluation of the expected effects of these changes on the function is relatively straightforward. In case various sequence homologs can be collected using BLAST, the sequences can be aligned and the conservation of every residue in the sequence can be evaluated (see Figure 10). Changes in the conserved residues (low natural variability) can be expected to affect the general function whereas changes in highly variable residues can be expected to cause a limited effect (Bharatham, Zhang, and Mihalek 2011). Moreover, when the group of homologs is divided in to its constituent groups of orthologs, the differences in conservation patterns between the various groups can be used to predict those changes that could affect the specific function (Kumar, Henikoff, and Ng 2009). In case structural data are available for one of the homologs the assessment of potential effects can be refined through homology modeling (Venselaar et al. 2010b; Worth, Preissner, and Blundell 2011) (see Figure 14).

b) Function prediction of a single sequence

For many sequences that have not received a specific function annotation in the past due to a lack of evidence in the experimental literature, orthologs or homologs can be found that have been related to such evidence. One could therefore routinely perform the initial steps of a sequence-based annotation as depicted in Figure 9 (and Appendix A). Moreover, as the number of sequenced genomes and therewith the number of potential orthologs and homologs rapidly increases, and likewise the volume of enzyme specific experimental data, it is often rewarding to follow this procedure.

c) Behavior of a single sequence in a different host

Another kind of modification involving sequences of known function might be the introduction in a different organism than the parental organism. One way to evaluate the potential effects of such a modification involves the reconstruction of the metabolism of the recipient organism with the new functionality. It might be that certain metabolic conversions that were not possible before could have become possible. Likewise, the presence of new promiscuous functions could have a similar effect.

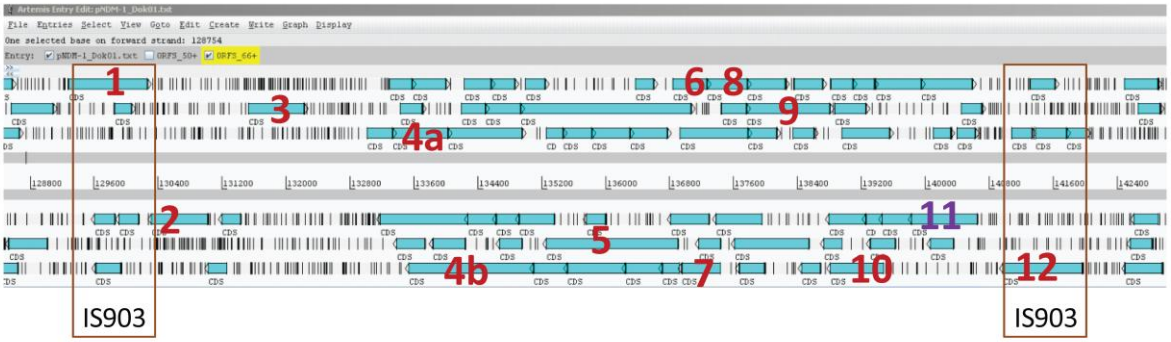
d) Function prediction of multiple sequences or multiple domains

The comparative analysis of a protein sequence that consists of multiple domains is not very different from the comparative analysis of multiple sequences. Basically, the individual domains or sequences should be used to search for homologs and orthologs and the results can then be treated as in the single sequence case. For sequences that lack clear boundaries, one could resort to the analysis of constituting partial sequences (e.g. 50 amino acids). Similarly, in the case the genuine ORFs can not be determined easily on basis of the DNA sequence one could take all potential ORFs (six reading frames) and perform a comparative analysis. This procedure is depicted in Figure 15 and Appendix B.

II) VARIATION IN DNA AND RNA SEQUENCE ELEMENTS

The evaluation of the potential effect of nucleotide changes for the 'non-coding' part of a DNA sequence is much harder than for the 'coding' part. This is the consequence of two interrelated phenomena namely: the relative lack of comprehensive knowledge concerning functional elements on the DNA (including for instance protein binding-sites or elements that encode small RNAs), and the fact that DNA is composed of only four different building blocks making the recognition of patterns more difficult than in the case of proteins (4 components vs. 20). With regard of the latter, the overall DNA structure is also far more homogeneous and less dependent on composition than protein structure. Nevertheless, some general rules do apply. And these can be used to interpret the potential function of particular DNA sequences as illustrated in Figure 16 and further discussed below.

A



B

1) IS element 903 [0]
2) putative zeta toxin [E-8]
3) Transposon Tn21 resolvase [E-37]
4) a, hypothetical [E-109]; b, transposase [0]
5) GroEL [0]
7) GroES [E-40]
6) unnamed [E-8]
8) cutA1 [E-20]
9) oxidoreductase domain protein [E-32]
10) trpCF gene product [E-25]
11) beta-lactamase [E-34]
12) IS element 903 [0]

Figure 15. Multiple BLAST analysis of the region of plasmid pNMD-1_Dok01 containing the *bla*$_{NMD-1}$ gene flanked by the IS903 elements

The nucleotide sequence of plasmid pNMD-1_Dok01 from *E. coli* was obtained from NCBI and loaded in to Artemis (Carver et al. 2008). **A)** Potential coding sequences were identified (>66 amino acids) in six reading frames for the region flanked by the IS elements 903. All sequences were BLASTed (cut-off E-value of 10) against the Enterobacterial genome sequences (data given in Appendix B). **B)** There were 13 sequences with a score <E-5. Of these only four overlap (i.e. 4a and 4b, 6 and 7). The identification matches perfectly with the one described by (Sekizuka et al. 2011). The plasmid contains the gene encoding NDM-1 metallo beta-lactamase, which imposes a high level of antibiotic resistance. The IS elements were also automatically identified using IS FInder (IBCG 2011) (see Appendix B).

In 2009 a carbapenem resistant *Klebsiella pneumoniae* strain was isolated from a Swedish patient that visited New Delhi (Yong et al. 2009). It appeared that the resistance was brought about by a new metallo-beta-lactamase encoded by the *blaNMD-1* gene which was located on a plasmid flanked by two IS elements. The metallo-beta-lactamases confer resistance to almost all beta-lactams (the major group of antibiotics) (Cornaglia, Giamarellou, and Rossolini 2011). Moreover, the *blaNMD-1* gene has proven extremely mobile between various bacterial species (Walsh et al. 2011). The mobility of the antibiotics-resistance conferring metallo-beta-lactamase among Gram-negative bacteria is currently one of the most important challenges in the control of infectious disease (Boucher et al. 2009).

a) Protein binding-sites

All genes (or better, units that are to be transcribed) are preceded by a promoter to recruit the RNA polymerase and often also by one or more binding-sites for regulatory proteins or regulatory RNA. In the case of regulatory proteins many DNA-sequence binding motifs have been defined on the basis of experimental evidence and many of these have been accumulated in reference databases (see Table 14). The defined motifs can be used to search potential binding sites on a genome. However the correct identification of binding-sites is not always a simple exercise due to the fact that many motifs are degenerate, i.e. that they represent many possible sequences. It is not always true that a sequence that fits a degenerate motif could actually serve as a binding site. Therefore careful definition of a specific motif is essential for the recognition of 'true' sites (Francke et al. 2008). Fortunately, the composition of a binding site is constrained by the molecular nature of the binding process and the helical nature of the DNA molecule. Since

most regulator proteins interact with the DNA via a helix-turn-helix sequence and as a dimer, a binding-site will in general be made up of two monomer binding sites and will either be palindromic or represent a direct repeat. Moreover, since the DNA is helical the monomer binding-site in general is between 5 and 7 nucleotides long and the two sites that make up the dimer binding-site are interspaced by a fixed number of nucleotides (Figure 16).

In general regulatory proteins belong to larger protein families and thus often more homologs are present per genome. The members of a regulator-protein family will in general adopt the same fold and as a result the DNA-binding motif should be similar (i.e. similar composition, and the same size and spacing), but still different enough to ensure regulator specificity[8]. As a consequence single or double nucleotide changes in a binding site can change the responsiveness to a different regulator of the same regulator family. A change in the nucleotides that characterize the family often causes the abolition of regulation. Likewise, small changes in the helix-turn-helix sequence of the regulator might cause a shift in regulon. The above is illustrated by the example given in Figure 17. Similar considerations hold for the analysis of eukaryotic sequences although TF-selectivity is complicated by the increased genome size which alters the probability of finding a particular binding sequence on the genome by chance considerably and thus, by necessity, additional factors must play a role (Pan et al. 2010). In case one does not know whether a sequence contains particular protein binding-sites it is very hard to recognize them because of the fact that the associated palindromes or direct-repeats are never perfect and they are short (discussed in (Nain, Sahi, and Kumar 2011)). A final factor to take in to account is the spacing of the binding-site with respect to the promoter. To achieve activation the binding site has to be located just upstream (~6 nucleotides) of the distant promoter element. However, repression can be achieved by binding on top of the promoter, but also downstream even extending into the coding sequence.

b) Structural elements on the RNA

Another main group of regulatory elements is related to forming structured RNA (Wan et al. 2011). Structure forming elements can be recognized on basis of the fact that they should contain palindromic sequences. The constituent parts (at least 2) may be interspaced, but they have to perfectly match[9] (in contrast to palindromic sequences related to transcription factor binding). The known RNA-elements vary widely in number, related to the kind of molecules they are associated with. Structural elements that occur in high numbers and/or contain longer conserved sequence stretches (>12 in prokaryotic genomes (see 8)) can be recognized relatively easily. Structural elements that are linked to a single inducer molecule mostly occur one or a few times in a genome. Finally, the position of the end of the element with respect to the translation start is rather invariable (again in contrast to transcription factor binding sites). For instance, riboswitches are structure forming regulatory elements at the 5' untranslated region of an mRNA molecule that can change conformation depending on the binding of an effector molecule, mostly a metabolite (Vitreschak et al. 2004; Tucker and Breaker 2005; Breaker 2011). The conformational change can terminate transcription (via the formation of a terminator structure) or allow read through (via the formation of an anti-terminator structure). A very similar mechanism is represented by the T-box, which is encountered in the control of amino acid transport and metabolism in low GC gram positive bacteria (Vitreschak et al. 2008; Wels et al. 2008; Gutierrez-Preciado et al. 2009)).

8 The chance of finding a particular nucleotide sequence of length 10 by chance is 4^{10} ≈1in 1.1 Mbase. As most prokaryotic genomes range in size from 2-5 Mbase, the number of retrieved correct 10 nucleotide sequences by chance is ~2-5. For a 12 nucleotide sequence the number becomes smaller than 1. In the human or a plant genome the number of nucleotides needed to reach a random probablity <1 is about 5 nucleotides longer.

9 One has to keep in mind that a G can pair with a U in mRNA.

A

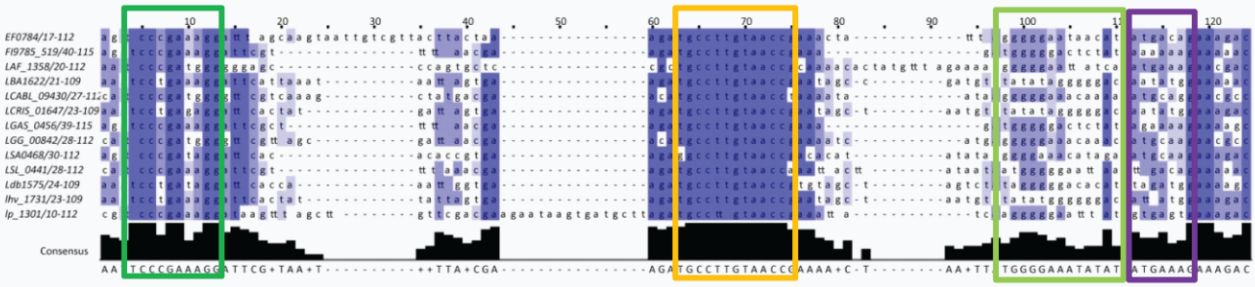
TF Lp_0172 from *Lactobacillus plantarum*

Motif:

match seq	score	relative orientation	distance	gene	product
GCAATCGCTTGCA	1	>>	61	lp_0172	TF
GCAAACGCTTGCA	1	>>	71	lp_1729	transport
GCAACCGGTTGCA	0.96	<<	12	lp_0064	hypothetical
GCAACCGGTTGCA	0.96	>>	158	lp_0066	beta-pgm
GCAATCGCTTACA	0.96	<<	357	lp_2643	lipoate-protein ligase
GCAAACGCTTACA	0.96	<<	303	lp_3316	hypothetical
GCAAACGCTTACA	0.96	><	242	lp_3590	hypothetical
GCAAACAGTTGCA	0.91	>>	-836	pWCFS101_01	replication
ACAATCGGTTCCA	0.91	><	-361	pWCFS103_30	hypothetical
TCAAACGGTTGCA	0.91	>>	811	lp_0025	alpha-amylase

B

The S_{MK}-box in various Lactobacilli



C

IS904 from *Lactococcus lactis*

Motif: TGGAAAG- -AACTAGACACA- -GTTAAGAGAA

match	score	start	upstream
AACTAGACACA	1.0000	53900	tra904A
AACTAGACACA	1.0000	137799	tra904B
AACTAGACACA	1.0000	140045	tra904C
AACTAGACACA	1.0000	373520	tra904D
AACTAGACACA	1.0000	626561	tra904E
AACTAGACACA	1.0000	835893	tra904F
AACTAGACACA	1.0000	838947	tra904G
AACTAGACACA	1.0000	2154611	tra904H
AACTAGACACA	1.0000	2214945	tra904I

Figure 16 Identification of regulatory and other DNA-sequence elements.

A) Identification of putative binding-sites for the transcription factor Lp_0172 in the *L. plantarum* genome. The binding motif was determined using a comprehensive footprinting approach set out in (Francke et al. 2008) and the motif was searched using a similar motif search procedure (Francke et al. 2011). The results are characteristic for most searches of potential transcription factor binding sites. Several putative sites were recovered with slightly varying sequence and with a varying orientation (indicated by arrows) and distance to the translation start of the gene downstream. Indicated in green are the sites that are probably genuine binding-sites, in blue the sites that could be binding sites (correct orientation) although the distance to the translation start is rather large, and in red those sites that are probably not

B) The structure forming S_{MK}-box upstream of *metK* in various *Lactobacillus* genomes. The translation start is positioned in the purple box and the first pairing element is found just upstream (in green), the second element of the pair is found at variable spacing around 50 nucleotides upstream (in green). In-between a conserved sequence that could represent a different regulatory element (in orange) is located. The S_{MK}-box is not easily recognized by eye. However, it is perfectly recognized using a motif search because of the perfect pairing.

C) Identification of IS element 904 in the genome of *L. lactis IL1403*. IS element 904 was found flanking a transposon encoding e.g. nisin production (Rauch, Beerthuyzen, and de Vos 1990). IS elements can be relatively easily recognized as they are longer than transcription factor binding sites, as they come in pairs flanking a transposon, and as there are mostly several perfect copies found in a particular genome.

In the case of the T-box, a terminator structure is formed shortly after transcription initiation unless an uncharged tRNA related to a specific amino acid binds to the specifier codon present in the T-box element, whereupon the anti-terminator structure is formed. The T-box element is around 200-250 nucleotides long and, as it contains four relatively long conserved sequence stretches, is easily recognizable on the genome. A change in the conserved parts does greatly affect the functioning of the element, whereas a change in the non-conserved part does not. Nevertheless, the substrate specificity is encoded in a single triplet, which is not always easily recognized from the multiple sequence alignment (illustrated in Figure 17). Other structural elements are far less easy to recognize because they involve less conserved residues. For example, the SMK-box found in low-GC gram positive bacteria is characterized by two conserved short stretches of 6-7 nucleotides (that are palindromic because they have to pair), which can be interspaced by a stretch of 40-200 nucleotides (Fuchs, Grundy, and Henkin 2006) (Figure 16). The inducing substrate is always S-adenosylmethionine (SAM).

c) Structural elements on the DNA

A third, more diverse group of nucleotide sequence elements relates more directly to the DNA itself. They include elements that impose structure-forming capabilities on the DNA like recombination hotspots (Smith 1994; Mezard 2006) or the REPs (BIMEs) found in the enterobacteria (Bachellier, Clement, and Hofnung 1999). In addition they include elements that enable the mobility of DNA sequence like CRISPRs (Horvath and Barrangou 2010; Al-Attar et al. 2011), which are related to the bacterial viral defense system, but also transposable elements (Beauregard, Curcio, and Belfort 2008). The latter elements are widely being exploited for engineering purposes (Ivics and Izsvak 2010; Palazzoli et al. 2010) because they allow directed integration of 'foreign' genetic material. Although much research has been devoted to the mobility of sequence elements it has until now proven difficult to predict without *a priori* knowledge whether a certain element confers this property and what the effect will be of a particular sequence change. Conversely, with *a priori* knowledge the task becomes easier. For instance, as the sequence of the IS elements associated with a particular transposon are completely conserved they are easily tractable within a given genome (see Figure 16). This property is for instance exploited by the IS Finder resource of CNRS, the LMGM at the University of Toulouse and UC Louvain-la-Neuve (IBCG 2011).

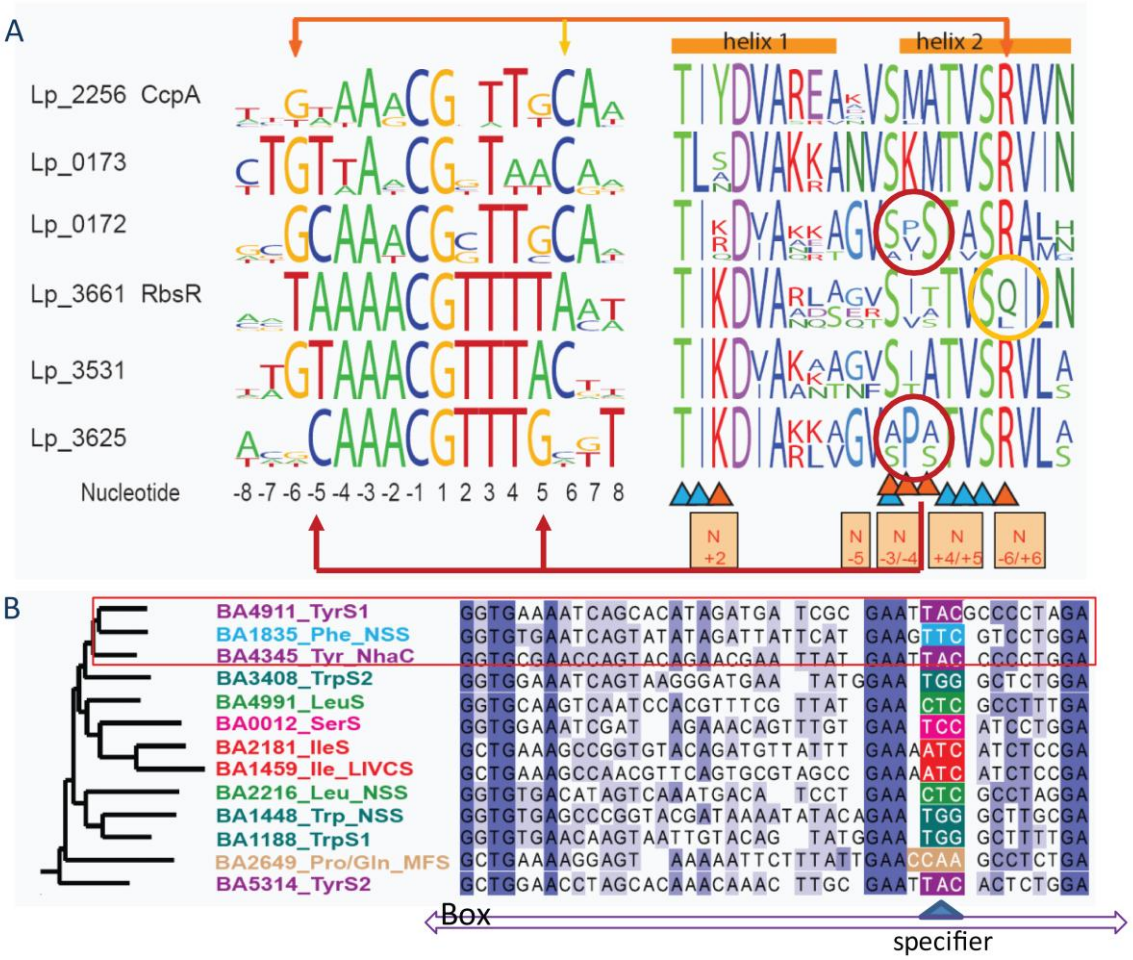


Figure 17. The natural variability of nucleotide sequence and its relation to function.
Similar to protein sequences one may assume for a set of orthologous/homologous nucleotide sequences, that those residues that have been conserved contribute more to the molecular function than those that have not been conserved. Vice versa one may assume that a change in the former will have an effect on function, whereas a change in the latter will not. Moreover a distinction can be made between those nucleotides that are conserved between all compared sequences, i.e. which will mostly confer general functional characteristics, and residues that are conserved between a subset of sequences, i.e. which will confer more specific characteristics. **A)** Correlation in the conservation of the transcription factor (TF) sequence and the putative binding site on the DNA for LacI-family proteins in *L. plantarum*. The picture projects the sequence conservation of the DNA-binding Helix Turn Helix motif within a particular orthologous group of the LacI-family of TFs next to the sequence conservation of the predicted binding site (derived from figure 2 in (Francke et al. 2008)). A clear correlation can be observed between the composition of the binding site at positions -6, -5 and +5, +6 and the change in conserved residues in helix 2 between the TF proteins. This pattern corresponds well with the crystal structure data (interaction indicated in boxes). Given these observations one would predict that changing the conserved Arg (R) of helix 2 of transcription factor Lp_3531 in to a Gln (Q) would affect the binding affinity to the original Lp_3531 binding site and would confer affinity to the binding sites of TF LP_3661, thereby making the related regulon sensitive to a different stimulus. At the same time one would predict almost no effect of a change in the Asp (D) of helix 1 in TF Lp_3531. **B)** Variability in the composition of the T-box element within the *Bacillus anthracis* genome. The data were taken from (Wels et al. 2008) and relates to Figure 12. The picture displays the part of the T-box sequence that contains the specifier codon. On the left a putative evolutionary lineage of the elements is given together with the annotation of the gene downstream. Several nucleotides are conserved between all T-box elements. Given the presented data one should expect that a change in any of those could cause the element to become disfunctional. On the other hand a single nucleotide change in the specifier codon will cause the element to become responsive to a different amino acid. In fact, we and others (Vitreschak et al. 2008; Wels et al. 2008) have hypothesized that this is actually the way in which the element evolves (see red box and Figure 12).

4) CONCLUSIONS AND FUTURE

The availability of whole genome sequence information has caused a considerable expansion in the range of biotechnological applications. Many of the new applications pose a challenge to the assessment of the associated risk. One of the main issues to be dealt with relates to the function annotation of extensively modified sequences and, similarly, to the annotation of 'new' sequences. In this report we illustrate that comparative genomics approaches are well-suited to support the identification of the potential biological function of a modified sequence. The main driver for the recognition of a functional relation is the consideration of evolutionary conservation. For instance, a sequence can be annotated specifically in case it is orthologous to, or has a similar genomic context to, a sequence with experimentally validated function. In case the evolutionary relations are less clear a prediction of function becomes less straightforward. Nevertheless, often a more generalized annotation can be achieved on the basis of family characteristics of the sequence (e.g. via profiles). By putting the function of the modified sequence in to the context of the reconstructed metabolism (or regulatory network) of a host organism, the potential for new biological roles of the modified sequence can be evaluated.

As the tools and pipelines needed to search (BLAST) and compare sequences (e.g. via multiple alignments or homology modeling) and to reconstruct metabolism are readily available and work well, the annotation procedure described in the above can to a large extent be standardized. In fact, the standardized use of BLAST searches in the assessment procedure for modified sequences was hinted at by the European Food Safety Authority (EFSA 2011) and was recently included in the US government federal guidance for DNA-synthesizing companies to assess the potential harmfulness of products of functions of synthetic sequences (US-government 2010). A practical bioinformatic implementation of the guidance is given in (Adam et al. 2011). It includes serial BLAST searches using a sliding window of around 200bp (in all 6 reading frames). This procedure in principle represents a good way to find potential culprits in most modified gene/protein sequences. However, the procedure will currently not be very effective to identify functional RNA- or DNA-sequence elements. The formulation and implementation of standardized analysis procedures that support the identification of these elements, and therewith support the reconstruction of regulatory networks, will require a substantial amount of additional research.

An aspect that is hardly discussed in this report, but which will become increasingly important, is the use of high throughput data to a priori substantiate assertions concerning the effect of a particular modification. One way to use the data would be to predict the functional context of the modified sequence given a particular environment. However, at this moment the data, like provided by microarrays (e.g. (Tengs et al. 2007)), RNA-seq (e.g. (Tengs et al. 2009)), or proteome (e.g. (Anttonen et al. 2010)) and metabolome (e.g. (Harrigan, Martino-Catt, and Glenn 2007)) are mainly being used to check a posteriori the positive and adverse effects of a modification.

Another aspect that is only slightly touched upon in this report involves the management of function information. The central administration of how particular sequences, organisms and environments relate to particular hazards is essential to support evaluations/claims made using (semi-)automated assessment procedures. The stored information will also be useful in the monitoring phase. In fact, appropriate handling and storage of the data obtained from monitoring should be(come) an important part of the monitoring strategy as with the *a posteriori* information (after the modification), *a priori* (before the modification) inferences can be checked and improved.

One of the main problems associated to bioinformatics-driven analysis is an overflow of information. The amount of data that can be accumulated and the variability within the data, make that sifting the relevant information from the noise in a reasonable amount of time is not always trivial. In this respect a clearly defined sifting procedure together with an integrated presentation of the data is an important aspect of an effective analysis. For example in the case of sequence searches, a comprehensive visualiza-

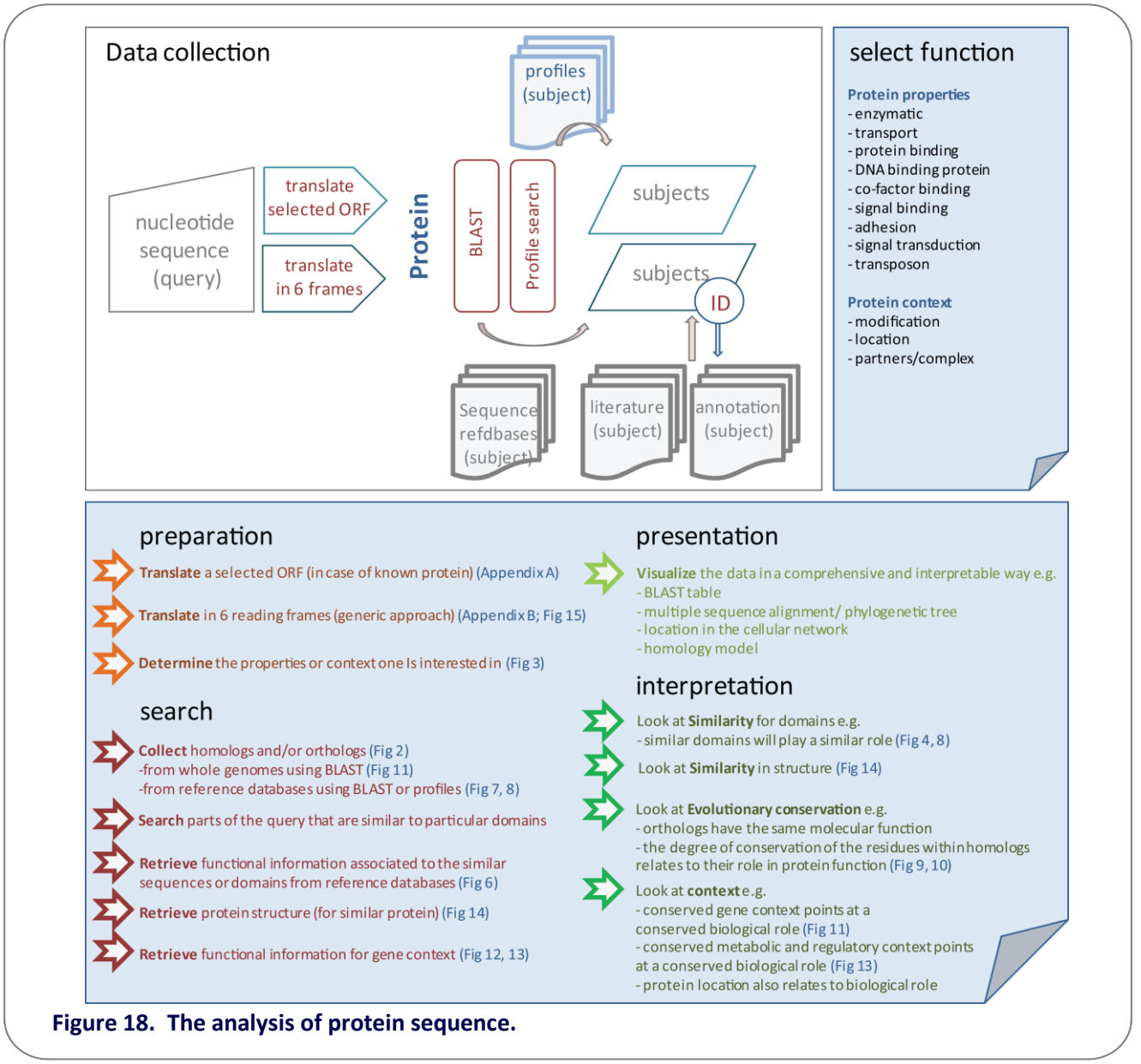


Figure 18. The analysis of protein sequence.

tion can not only alleviate the burden of interpreting all of the BLAST results in a reasonable amount of time, it can also prove essential for a correct interpretation of the results (e.g. in case of overlapping hits).

One of the cornerstones of the current risk assessment strategies is the consideration related to a history of safe use. This concept is powerful in the sense that many conditions have been probed in the past and the absence of any negative report then makes a claim towards safe use of a certain application or organism more trustworthy. The former is of course only true when probing has occurred under conditions where safety of use has been challenged. In a similar way the history of a sequence is the cornerstone of sequence-based predictions. The selective pressures experienced in the past to maintain a particular function have induced particular patterns of residue conservation in any given sequence. Thus the degree of conservation between evolutionary related sequences may be related to the importance of a certain residue for the particular function that has been conserved. And a change in such a residue may be expected to affect the function. On the other hand, the replacement of a residue or residues that is/are hardly conserved (i.e. show a high natural variability) may be expected to induce little effect.

The variability in sequence between different organisms has been exploited in the estimation of for instance plasmid promiscuity (Suzuki et al. 2010) or the extent of lateral gene transfer (Langille and Brinkman 2009). Similarly, the

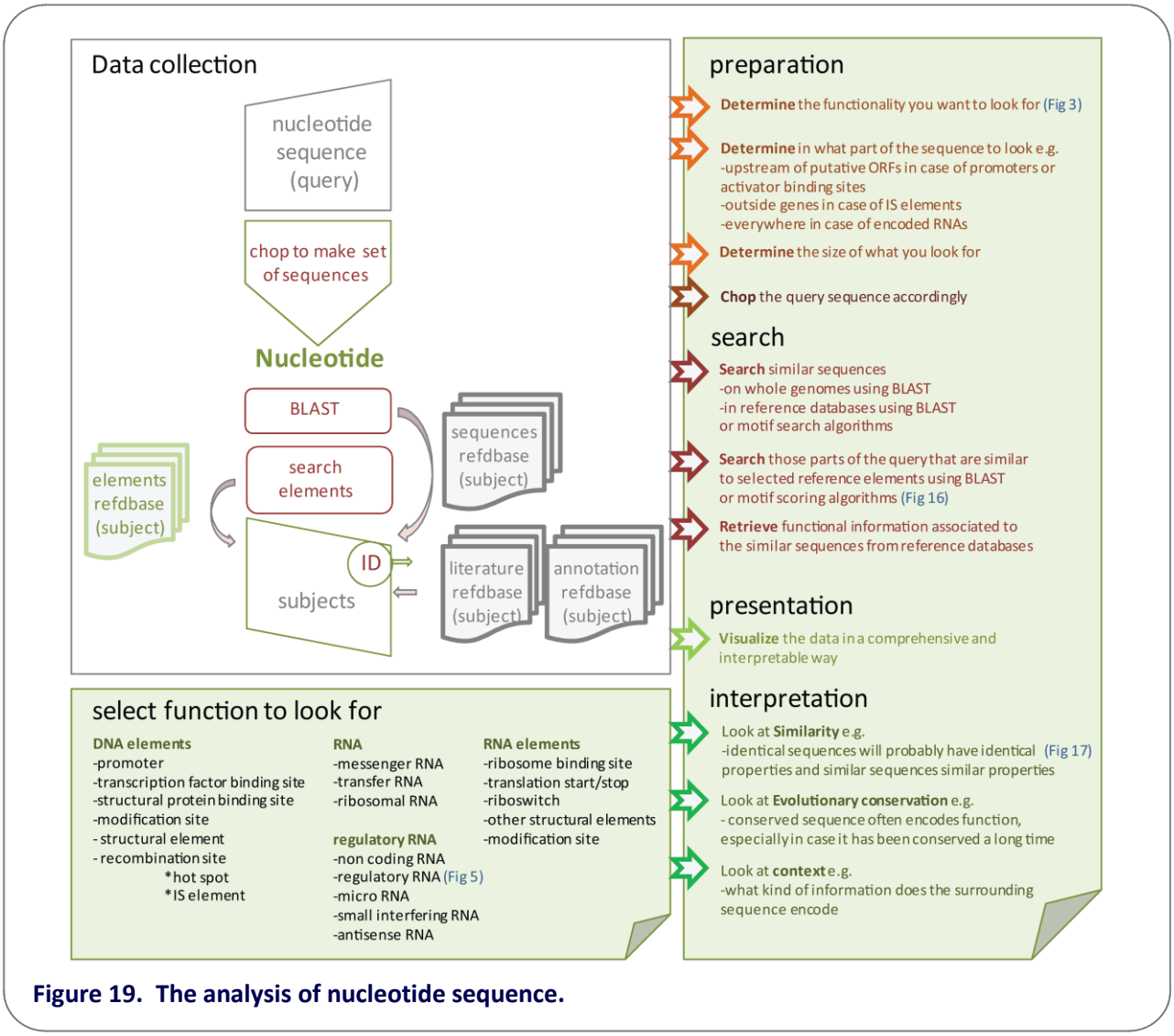


Figure 19. The analysis of nucleotide sequence.

transferability of sequence material can be calculated from a statistical analysis of the compositional variation within a population, like done for crops by (Harrigan et al. 2010). Nevertheless, currently these estimates lack the accuracy to calculate the probability that a certain sequence can/will be transferred. Considering the ongoing research in this field it is not unlikely that in the future one can come up with relatively reliable estimates for the probability of sequence transfer events.

Whatever the biotechnological application, the assessment of risk will be restricted with respect to the available time by necessity. Because the assessment is comparative by nature, comparative genomics approaches fit well in to the current practice. The added value of the application of bioinformatics analysis techniques predominantly relates to the first phase of the assessment: the identification of the potential consequences (i.e. implied hazard) of certain sequence modifications. Although the related methodology has become standardized, one should be aware that the interpretation of the results in many cases requires experienced eyes (Jones and Swindells 2002). The proposed assessment process can be summarized as depicted in Figures 18 and 19, for the identification of protein and nucleotide function, respectively. First, the kind of molecular/ physiological properties one is interested in have to be defined. Then the relevant data can be collected from the appropriate sources of information. And third, the data should then be presented in a comprehensive way to enable a quick interpretation. Finally, the increasing amount of data and concurrent development of tools will increase the predictive power of sequence analyses and thereby will move the value of bioinformatics also in the realm of evaluating risk in more quantitative terms.

5) REFERENCES

Abouheif, E., M. Akam, W. J. Dickinson, P. W. Holland, A. Meyer, N. H. Patel, R. A. Raff, V. L. Roth, and G. A. Wray. 1997. Homology and developmental genes. Trends Genet 13:432-433.

Adam, L., M. Kozar, G. Letort, O. Mirat, A. Srivastava, T. Stewart, M. L. Wilson, and J. Peccoud. 2011. Strengths and limitations of the federal guidance on synthetic DNA. Nat Biotechnol 29:208-210.

Agrawal, A., and X. Huang. 2011. Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices. IEEE/ACM Trans Comput Biol Bioinform 8:194-205.

Al-Attar, S., E. R. Westra, J. van der Oost, and S. J. Brouns. 2011. Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes. Biol Chem 392:277-289.

Aleshin, A. E., P. H. Feng, R. B. Honzatko, and P. J. Reilly. 2003. Crystal structure and evolution of a prokaryotic glucoamylase. J Mol Biol 327:61-73.

Alkema, W. B., B. Lenhard, and W. W. Wasserman. 2004. Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. Genome Res 14:1362-1373.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol 215:403-410.

Altschul, S. F., and E. V. Koonin. 1998. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem Sci 23:444-447.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Andersen, J. M., R. Barrangou, M. Abou Hachem, S. Lahtinen, Y. J. Goh, B. Svensson, and T. R. Klaenhammer. 2011. Transcriptional and functional analysis of galactooligosaccharide uptake by lacS in Lactobacillus acidophilus. Proc Natl Acad Sci U S A 108:17785-17790.

Andersson, U., F. Levander, and P. Radstrom. 2001. Trehalose-6-phosphate phosphorylase is part of a novel metabolic pathway for trehalose utilization in Lactococcus lactis. J Biol Chem 276:42707-42713.

Andreeva, A., D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36:D419-425.

Aniba, M. R., O. Poch, and J. D. Thompson. 2010. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Res 38:7353-7363.

Anttonen, M. J., S. Lehesranta, S. Auriola, R. M. Rohlig, K. H. Engel, and S. O. Karenlampi. 2010. Genetic and environmental influence on maize kernel proteome. J Proteome Res 9:6160-6168.

Arai, M., K. Okumura, M. Satake, and T. Shimizu. 2004. Proteome-wide functional classification and identification of prokaryotic transmembrane proteins by transmembrane topology similarity comparison. Protein Sci 13:2170-2183.

Arnaoudova, E., D. C. Haws, P. Huggins, J. W. Jaromczyk, N. Moore, C. L. Schardl, and R. Yoshida. 2010. Statistical phylogenetic tree analysis using differences of means. Front Neurosci 4.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25-29.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Bachellier, S., J. M. Clement, and M. Hofnung. 1999. Short palindromic repetitive DNA elements in enterobacteria: a survey. Res Microbiol 150:627-639.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202-208.

Bairoch, A. 2000. The ENZYME database in 2000. Nucleic Acids Res 28:304-305.

Baldwin, T. K., R. Winnenburg, M. Urban, C. Rawlings, J. Koehler, and K. E. Hammond-Kosack. 2006. The pathogen-host interactions database (PHI-base) provides insights into generic and novel themes of pathogenicity. Mol Plant Microbe Interact 19:1451-1462.

Baran, J., M. Gerner, M. Haeussler, G. Nenadic, and C. M. Bergman. 2011. pubmed2ensembl: a resource for mining the biological literature on genes. PLoS One 6:e24716.

Bayer, T. S. 2010. Grand challenge commentary: Transforming biosynthesis into an information science. Nat Chem Biol 6:859-861.

Beauregard, A., M. J. Curcio, and M. Belfort. 2008. The take and give between retrotransposable elements and their hosts. Annu Rev Genet 42:587-617.

Beisvag, V., F. K. Junge, H. Bergum, L. Jolsum, S. Lydersen, C. C. Gunther, H. Ramampiaro, M. Langaas, A. K. Sandvik, and A. Laegreid. 2006. GeneTools--application for functional annotation and statistical hypothesis testing. BMC Bioinformatics 7:470.

Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783-795.

Benson, D. A., I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E. W. Sayers. 2012. GenBank. Nucleic Acids Res 40:D48-53.

Bergmans, H., C. Logie, K. Van Maanen, H. Hermsen, M. Meredyth, and C. Van Der Vlugt. 2008. Identification of potentially hazardous human gene products in GMO risk assessment. Environ Biosafety Res 7:1-9.

Bhagwat, and Aravind. 2007. PSI BLAST tutorial. NCBI Bookshelf.

Bharatham, K., Z. H. Zhang, and I. Mihalek. 2011. Determinants, discriminants, conserved residues--a heuristic approach to detection of functional divergence in protein families. PLoS One 6:e24382.

Blackshields, G., I. M. Wallace, M. Larkin, and D. G. Higgins. 2006. Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biol 6:321-339.

Blazeck, J., and H. Alper. 2010. Systems metabolic engineering: genome-scale models and beyond. Biotechnol J 5:647-659.

Bordbar, A., and B. O. Palsson. 2011. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. J Intern Med 271:131-141.

Bork, P., T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan. 1998. Predicting function: from genes to genomes and back. J Mol Biol 283:707-725.

Born, T. L., and J. S. Blanchard. 1999. Enzyme-catalyzed acylation of homoserine: mechanistic characterization of the Escherichia coli metA-encoded homoserine transsuccinylase. Biochemistry 38:14416-14423.

Bornberg-Bauer, E., A. K. Huylmans, and T. Sikosek. 2010. How do new proteins arise? Curr Opin Struct Biol 20:390-396.

Boucher, H. W., G. H. Talbot, J. S. Bradley, J. E. Edwards, D. Gilbert, L. B. Rice, M. Scheld, B. Spellberg, and J. Bartlett. 2009. Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. Clin Infect Dis 48:1-12.

Bouige, P., D. Laurent, L. Piloyan, and E. Dassa. 2002. Phylogenetic and functional classification of ATP-binding cassette (ABC) systems. Curr Protein Pept Sci 3:541-559.

Boyle, P. M., and P. A. Silver. 2011. Parts plus pipes: Synthetic biology approaches to metabolic engineering. Metab Eng.

Bradshaw, C. R., V. Surendranath, and B. Habermann. 2006. ProFAT: a web-based tool for the functional annotation of protein sequences. BMC Bioinformatics 7:466.

Brahmachary, M., S. P. Krishnan, J. L. Koh, A. M. Khan, S. H. Seah, T. W. Tan, V. Brusic, and V. B. Bajic. 2004. ANTIMIC: a database of antimicrobial sequences. Nucleic Acids Res 32:D586-589.

Brandt, B. W., and J. Heringa. 2011. Protein analysis tools and services at IBIVU. J Integr Bioinform 8:168.

Breaker, R. R. 2011. Prospects for riboswitch discovery and analysis. Mol Cell 43:867-879.

Brenner, S. E. 1999. Errors in genome annotation. Trends Genet 15:132-133.

Bru, C., E. Courcelle, S. Carrere, Y. Beausse, S. Dalmar, and D. Kahn. 2005. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 33:D212-215.

Bryson, K., L. J. McGuffin, R. L. Marsden, J. J. Ward, J. S. Sodhi, and D. T. Jones. 2005. Protein structure prediction servers at University College London. Nucleic Acids Res 33:W36-38.

Bucher, P., K. Karplus, N. Moeri, and K. Hofmann. 1996. A flexible motif search technique based on generalized profiles. Comput Chem 20:3-23.

Bulyk, M. L. 2003. Computational prediction of transcription-factor binding site locations. Genome Biol 5:201.

Burgetz, I. J., S. Shariff, A. Pang, and E. R. Tillier. 2006. Positional homology in bacterial genomes. Evol Bioinform Online 2:77-90.

Caboche, S., M. Pupin, V. Leclere, A. Fontaine, P. Jacques, and G. Kucherov. 2008. NORINE: a database of nonribosomal peptides. Nucleic Acids Res 36:D326-331.

Campbell, J. H., J. A. Lengyel, and J. Langridge. 1973. Evolution of a second gene for beta-galactosidase in Escherichia coli. Proc Natl Acad Sci U S A 70:1841-1845.

Cantarel, B. L., P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37:D233-238.

Carbon, S., A. Ireland, C. J. Mungall, S. Shu, B. Marshall, and S. Lewis. 2009. AmiGO: online access to ontology and annotation data. Bioinformatics 25:288-289.

Carbonell, P., and J. L. Faulon. 2010. Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics 26:2012-2019.

CARD. 2011. arpcard.mcmaster.ca. The Comprehensive Antibiotic Resistance Database Pilot Project. McMaster University, Canada.

Carr, P. A., and G. M. Church. 2009. Genome engineering. Nat Biotechnol 27:1151-1162.

Carver, T., M. Berriman, A. Tivey, C. Patel, U. Bohme, B. G. Barrell, J. Parkhill, and M. A. Rajandream. 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics 24:2672-2676.

Caspi, R., T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, and P. D. Karp. 2009. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 38:D473-479.

Cerami, E. G., B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. 2011. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res 39:D685-690.

Chan, P. P., A. D. Holmes, A. M. Smith, D. Tran, and T. M. Lowe. 2012. The UCSC Archaeal Genome Browser: 2012 update. Nucleic Acids Res 40:D646-652.

Chan, P. P., A. D. Holmes, A. M. Smith, D. Tran, and T. M. Lowe. 2011. The UCSC Archaeal Genome Browser: 2012 update. Nucleic Acids Res 40:D646-652.

Chandler, S. F., and F. Brugliera. 2011. Genetic modification in floriculture. Biotechnol Lett 33:207-214.

Chandler, S. F., and Y. Tanaka. 2010. Genetic modification in floriculture. Crit Rev Plant Sci 26:169-197.

Chen, N., I. J. Val, S. Kyriakopoulos, K. M. Polizzi, and C. Kontoravdi. 2011. Metabolic network reconstruction: advances in in silico interpretation of analytical information. Curr Opin Biotechnol.

Chen, X., L. Guo, Z. Fan, and T. Jiang. 2008. W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. Bioinformatics 24:1121-1128.

Clare, A., and R. D. King. 2002. Machine learning of functional class from phenotype data. Bioinformatics 18:160-166.

Claudel-Renard, C., C. Chevalet, T. Faraut, and D. Kahn. 2003. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res 31:6633-6639.

Conant, G. C., and K. H. Wolfe. 2008. GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics 24:861-862.

Cornaglia, G., H. Giamarellou, and G. M. Rossolini. 2011. Metallo-beta-lactamases: a last frontier for beta-lactams? Lancet Infect Dis 11:381-393.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188-1190.

Cserzo, M., F. Eisenhaber, B. Eisenhaber, and I. Simon. 2002. On filtering false positive transmembrane protein predictions. Protein Eng 15:745-52.

Culler, S. J., K. G. Hoff, and C. D. Smolke. 2010. Reprogramming cellular behavior with RNA controllers responsive to endogenous proteins. Science 330:1251-1255.

D'Eustachio, P. 2011. Reactome knowledgebase of human biological pathways and processes. Methods Mol Biol 694:49-61.

Dago, A. E., S. R. Wigneshweraraj, M. Buck, and E. Morett. 2007. A role for the conserved GAFTGA motif of AAA+ transcription activators in sensing promoter DNA conformation. J Biol Chem 282:1087-1097.

Danchin, A. 2004. The bag or the spindle: the cell factory at the time of systems' biology. Microb Cell Fact 3:13.

Das, S., M. P. Krein, and C. M. Breneman. 2010. PESDserv: a server for high-throughput comparison of protein binding site surfaces. Bioinformatics 26:1913-1914.

Dassa, E., and P. Bouige. 2001. The ABC of ABCS: a phylogenetic and functional classification of ABC systems in living organisms. Res Microbiol 152:211-229.

Datta, R. S., C. Meacham, B. Samad, C. Neyer, and K. Sjolander. 2009. Berkeley PHOG: PhyloFacts orthology group prediction web server. Nucleic Acids Res 37:W84-89.

Davidson, A. L., E. Dassa, C. Orelle, and J. Chen. 2008. Structure, function, and evolution of bacterial ATP-binding cassette systems. Microbiol Mol Biol Rev 72:317-364, table of contents.

Davies, K. M., E. D. Lowe, C. Venien-Bryan, and L. N. Johnson. 2009. The HupR receiver domain crystal structure in its nonphospho and inhibitory phospho states. J Mol Biol 385:51-64.

de Jong, A., A. J. van Heel, J. Kok, and O. P. Kuipers. 2010. BAGEL2: mining for bacteriocins in genomic data. Nucleic Acids Res 38:W647-651.

De Smet, R., and K. Marchal. 2010. Advantages and limitations of current network inference methods. Nat Rev Microbiol 8:717-729.

Dean, A. M. 2010. The future of molecular evolution. EMBO Rep 11:409.

Dehal, P. S., M. P. Joachimiak, M. N. Price, J. T. Bates, J. K. Baumohl, D. Chivian, G. D. Friedland, K. H. Huang, K. Keller, P. S. Novichkov, I. L. Dubchak, E. J. Alm, and A. P. Arkin. 2010. MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res 38:D396-400.

Deluca, T. F., J. Cui, J. Y. Jung, K. C. St Gabriel, and D. P. Wall. 2012. Roundup 2.0: Enabling comparative genomics for over 1800 genomes. Bioinformatics.

Dereeper, A., S. Audic, J. M. Claverie, and G. Blanc. 2010. BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BMC Evol Biol 10:8.

Derrien, T., C. Andre, F. Galibert, and C. Hitte. 2007. AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. Bioinformatics 23:498-499.

Deutscher, J., C. Francke, and P. W. Postma. 2006. How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. Microbiol Mol Biol Rev 70:939-1031.

Devos, D., and A. Valencia. 2000. Practical limits of function prediction. Proteins 41:98-107.

Devos, D., and A. Valencia. 2001. Intrinsic errors in genome annotation. Trends Genet 17:429-431.

Dewey, C. N. 2011. Positional orthology: putting genomic evolutionary relationships into context. Brief Bioinform 12:401-412.

Di Tommaso, P., S. Moretti, I. Xenarios, M. Orobitg, A. Montanyola, J. M. Chang, J. F. Taly, and C. Notredame. 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res 39:W13-17.

Dill, K. A., S. B. Ozkan, T. R. Weikl, J. D. Chodera, and V. A. Voelz. 2007. The protein folding problem: when will it be solved? Curr Opin Struct Biol 17:342-346.

Do, C. B., M. S. Mahabhashyam, M. Brudno, and S. Batzoglou. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15:330-340.

Duarte, N. C., S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. O. Palsson. 2007. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 104:1777-1782.

Duchrow, T., T. Shtatland, D. Guettler, M. Pivovarov, S. Kramer, and R. Weissleder. 2009. Enhancing navigation in biomedical databases by community voting and database-driven text classification. BMC Bioinformatics 10:317.

Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison. 1999. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids Cambridge University Press, Cambridge.

Durot, M., P. Y. Bourguignon, and V. Schachter. 2009. Genome-scale models of bacterial metabolism:

reconstruction and applications. FEMS Microbiol Rev 33:164-190.

EcARS. 2011. http://www.broadinstitute.org/. Escherichia coli Antibiotic Resistance Sequencing Project. Broad Institute of Harvard and MIT Boston.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Edwards, D., and J. Batley. 2010. Plant genome sequencing: applications for crop improvement. Plant Biotechnol J 8:2-9.

EFSA. 2006. Guidance document of the scientific panel on genetically modified organisms for the risk assessment of genetically modified plants and derived food and feed. EFSA Journal, 99:1-100.

EFSA. 2011. Guidance on the risk assessment of genetically modified microorganisms and their products intended for food and feed use. Pp. 2193. EFSA journal.

Egel, R. 2000. How 'homology' entered genetics. Trends Genet 16:437-439.

Egloff, M. P., J. Uppenberg, L. Haalck, and H. van Tilbeurgh. 2001. Crystal structure of maltose phosphorylase from Lactobacillus brevis: unexpected evolutionary relationship with glucoamylases. Structure 9:689-697.

Eisen, J. A., and M. Wu. 2002. Phylogenetic analysis and gene functional predictions: phylogenomics in action. Theor Popul Biol 61:481-487.

Eisenstein, M. 2010. Synthetic DNA firms embrace hazardous agents guidance but remain wary of automated 'best-match'. Nat Biotechnol 28:1225-1226.

Ellis, D. I., and R. Goodacre. 2011. Metabolomics-assisted synthetic biology. Curr Opin Biotechnol.

Erickson, B., R. Singh, and P. Winters. 2011. Synthetic biology: regulating industry uses of new biotechnologies. Science 333:1254-1256.

Falagas, M. E., and E. A. Karveli. 2006. World Wide Web resources on antimicrobial resistance. Clin Infect Dis 43:630-633.

Falth, M., K. Skold, M. Norrman, M. Svensson, D. Fenyo, and P. E. Andren. 2006. SwePep, a database designed for endogenous peptides and mass spectrometry. Mol Cell Proteomics 5:998-1005.

Farrar, M. 2007. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics 23:156-161.

Felsenstein, J. 2003. Inferring Phylogenies Sinauer Associates Inc., Sunderland, MA.

Felsenstein, J. 2004. cmgm.stanford.edu/phylip/. PHYLIP. Stanford University, Palo Alto.

Feng, D. F., and R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25:351-360.

Fernandez-Suarez, X. M., and M. K. Schuster. 2010. Using the ensembl genome server to browse genomic sequence data. Curr Protoc Bioinformatics Chapter 1:Unit1 15.

Fink, J. L., S. Kushch, P. R. Williams, and P. E. Bourne. 2008. BioLit: integrating biological literature with databases. Nucleic Acids Res 36:W385-389.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman. 2010. The Pfam protein families database. Nucleic Acids Res 38:D211-222.

Fischer, D. 2006. Servers for protein structure prediction. Curr Opin Struct Biol 16:178-182.

Fischer, S., B. P. Brunk, F. Chen, X. Gao, O. S. Harb, J. B. Iodice, D. Shanmugam, D. S. Roos, and C. J. Stoeckert, Jr. 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics Chapter 6:Unit 6 12 11-19.

Fisher, M. A., K. L. McKinley, L. H. Bradley, S. R. Viola, and M. H. Hecht. 2011. De novo designed proteins from a library of artificial sequences function in Escherichia coli and enable cell growth. PLoS One 6:e15364.

Fitch, W. M. 1970. Further improvements in the method of testing for evolutionary homology among proteins. J Mol Biol 49:1-14.

Fitch, W. M. 2000. Homology a personal view on some of the problems. Trends Genet 16:227-231.

Fleming, K., A. Muller, R. M. MacCallum, and M. J. Sternberg. 2004. 3D-GENOMICS: a database to compare structural and functional annotations of proteins between sequenced genomes. Nucleic Acids Res 32:D245-250.

Fong, C., L. Rohmer, M. Radey, M. Wasnick, and M. J. Brittnacher. 2008. PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes. BMC Bioinformatics 9:170.

Francke, C., T. Groot Kormelink, Y. Hagemeijer, L. Overmars, V. Sluijter, R. Moezelaar, and R. J. Siezen. 2011. Comparative analyses imply that the enigmatic Sigma factor 54 is a central controller of the bacterial exterior. BMC Genomics 12:385.

Francke, C., R. Kerkhoven, M. Wels, and R. J. Siezen. 2008. A generic approach to identify Transcription Factor-specific operator motifs; Inferences for LacI-family mediated regulation in Lactobacillus plantarum WCFS1. BMC Genomics 9:145.

Francke, C., R. J. Siezen, and B. Teusink. 2005. Reconstructing the metabolic network of a bacterium from its genome. Trends Microbiol 13:550-558.

Friedmann, H. C. 2004. From "butyribacterium" to "E. coli": an essay on unity in biochemistry. Perspect Biol Med 47:47-66.

Fröhlich, K. S., and J. Vogel. 2009. Activation of gene expression by small RNA. Curr Opin Microbiol 12:674-682.

Fuchs, R. T., F. J. Grundy, and T. M. Henkin. 2006. The S(MK) box is a new SAM-binding RNA for translational regulation of SAM synthetase. Nat Struct Mol Biol 13:226-233.

Fuellen, G. 2011. Evolution of gene regulation--on the road towards computational inferences. Brief Bioinform 12:122-131.

Gaeta, R. T., R. E. Masonbrink, L. Krishnaswamy, C. Zhao, and J. A. Birchler. 2011. Synthetic Chromosome Platforms in Plants. Annu Rev Plant Biol.

Gagliotti, C., A. Balode, F. Baquero, J. Degener, H. Grundmann, D. Gur, V. Jarlier, G. Kahlmeter, J. Monen, D. L. Monnet, G. M. Rossolini, C. Suetens, K. Weist, and O. Heuer. 2011. Escherichia coli and

Staphylococcus aureus: bad news and good news from the European Antimicrobial Resistance Surveillance Network (EARS-Net, formerly EARSS), 2002 to 2009. Euro Surveill 16.

Gama-Castro, S., H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muniz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. Garcia-Sotelo, A. Lopez-Fuentes, L. Porron-Sotelo, S. Alquicira-Hernandez, A. Medina-Rivera, I. Martinez-Flores, K. Alquicira-Hernandez, R. Martinez-Adame, C. Bonavides-Martinez, J. Miranda-Rios, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vides. 2011. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Res 39:D98-105.

Garcia-Remesal, M., A. Cuevas, D. Perez-Rey, L. Martin, A. Anguita, D. de la Iglesia, G. de la Calle, J. Crespo, and V. Maojo. 2010. PubDNA Finder: a web database linking full-text articles to sequences of nucleic acids. Bioinformatics 26:2801-2802.

Gardner, P. P., J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman. 2009. Rfam: updates to the RNA families database. Nucleic Acids Res 37:D136-140.

Gascuel, O., and M. Steel. 2006. Neighbor-joining revealed. Mol Biol Evol 23:1997-2000.

Gaudet, P., A. Bairoch, D. Field, S. A. Sansone, C. Taylor, T. K. Attwood, A. Bateman, J. A. Blake, C. J. Bult, J. M. Cherry, R. L. Chisholm, G. Cochrane, C. E. Cook, J. T. Eppig, M. Y. Galperin, R. Gentleman, C. A. Goble, T. Gojobori, J. M. Hancock, D. G. Howe, T. Imanishi, J. Kelso, D. Landsman, S. E. Lewis, I. K. Mizrachi, S. Orchard, B. F. Ouellette, S. Ranganathan, L. Richardson, P. Rocca-Serra, P. N. Schofield, D. Smedley, C. Southan, T. W. Tan, T. Tatusova, P. L. Whetzel, O. White, and C. Yamasaki. 2012. Towards BioDBcore: a community-defined information specification for biological databases. Nucleic Acids Res 39:D7-10.

Gel Moreno, B., A. M. Jenkinson, R. C. Jimenez, X. Messeguer Peypoch, and H. Hermjakob. 2011. easyDAS: automatic creation of DAS servers. BMC Bioinformatics 12:23.

George, R. A., and J. Heringa. 2002. Protein domain identification and improved sequence similarity searching using PSI-BLAST. Proteins 48:672-681.

Gerlt, J. A., and P. C. Babbitt. 2009. Enzyme (re)design: lessons from natural evolution and computation. Curr Opin Chem Biol 13:10-18.

Gladki, A., P. Siedlecki, S. Kaczanowski, and P. Zielenkiewicz. 2008. e-LiSe--an online tool for finding needles in the '(Medline) haystack'. Bioinformatics 24:1115-1117.

Glanville, J. G., D. Kirshner, N. Krishnamurthy, and K. Sjolander. 2007. Berkeley Phylogenomics Group web servers: resources for structural phylogenomic analysis. Nucleic Acids Res 35:W27-32.

Gollapudi, R., K. V. Revanna, C. Hemmerich, S. Schaack, and Q. Dong. 2008. BOV--a web-based BLAST output visualization tool. BMC Genomics 9:414.

Gonzalez-Diaz, H., F. Prado-Prado, E. Sobarzo-Sanchez, M. Haddad, S. Maurel Chevalley, A. Valentin, J. Quetin-Leclercq, M. A. Dea-Ayuela, M. Teresa Gomez-Munos, C. R. Munteanu, J. Jose Torres-Labandeira, X. Garcia-Mera, R. A. Tapia, and F. M. Ubeira. 2011. NL MIND-BEST: a web server for ligands and proteins discovery--theoretic-experimental study of proteins of Giardia lamblia and new compounds active against Plasmodium falciparum. J Theor Biol 276:229-249.

Gopel, Y., D. Luttmann, A. K. Heroven, B. Reichenbach, P. Dersch, and B. Gorke. 2011. Common and divergent features in transcriptional control of the homologous small RNAs GlmY and GlmZ in Enterobacteriaceae. Nucleic Acids Res 39:1294-1309.

Greene, L. H., T. E. Lewis, S. Addou, A. Cuff, T. Dallman, M. Dibley, O. Redfern, F. Pearl, R. Nambudiry, A. Reid, I. Sillitoe, C. Yeats, J. M. Thornton, and C. A. Orengo. 2007. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35:D291-297.

Gribskov, M. 1992. Translational initiation factors IF-1 and eIF-2 alpha share an RNA-binding motif with prokaryotic ribosomal protein S1 and polynucleotide phosphorylase. Gene 119:107-111.

Gribskov, M., A. D. McLachlan, and D. Eisenberg. 1987. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A 84:4355-4358.

Grote, A., J. Klein, I. Retter, I. Haddad, S. Behling, B. Bunk, I. Biegler, S. Yarmolinetz, D. Jahn, and R. Munch. 2009. PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. Nucleic Acids Res 37:D61-65.

Gueguen, Y., J. Garnier, L. Robert, M. P. Lefranc, I. Mougenot, J. de Lorgeril, M. Janech, P. S. Gross, G. W. Warr, B. Cuthbertson, M. A. Barracco, P. Bulet, A. Aumelas, Y. Yang, D. Bo, J. Xiang, A. Tassanakajon, D. Piquemal, and E. Bachere. 2006. PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. Dev Comp Immunol 30:283-288.

Guell, M., E. Yus, M. Lluch-Senar, and L. Serrano. 2011. Bacterial transcriptomics: what is beyond the RNA horiz-ome? Nat Rev Microbiol 9:658-669.

Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307-321.

Guo, J. T., K. Ellrott, and Y. Xu. 2008. A historical perspective of template-based protein structure prediction. Methods Mol Biol 413:3-42.

Gutierrez-Preciado, A., T. M. Henkin, F. J. Grundy, C. Yanofsky, and E. Merino. 2009. Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. Microbiol Mol Biol Rev 73:36-61.

Haeussler, M., M. Gerner, and C. M. Bergman. 2011. Annotating genes and genomes with DNA sequences extracted from biomedical articles.

Bioinformatics 27:980-986.
Hall, B. G. 2011. Phylogenetic Trees Made Easy: A How-To Manual. Sinauer Associates Inc., Sunderland, MA.
Hall, B. G. 2003. The EBG system of E. coli: origin and evolution of a novel beta-galactosidase for the metabolism of lactose. Genetica 118:143-156.
Hammami, R., J. Ben Hamida, G. Vergoten, and I. Fliss. 2009. PhytAMP: a database dedicated to antimicrobial plant peptides. Nucleic Acids Res 37:D963-968.
Hammami, R., A. Zouhir, C. Le Lay, J. Ben Hamida, and I. Fliss. 2010. BACTIBASE second release: a database and tool platform for bacteriocin characterization. BMC Microbiol 10:22.
Harrigan, G. G., D. Lundry, S. Drury, K. Berman, S. G. Riordan, M. A. Nemeth, W. P. Ridley, and K. C. Glenn. 2010. Natural variation in crop composition and the impact of transgenesis. Nat Biotechnol 28:402-404.
Harrigan, G. G., S. Martino-Catt, and K. C. Glenn. 2007. Metabolomics, metabolic diversity and genetic variation in crops. metabolomics 3:259-272.
He, J., X. Dai, and X. Zhao. 2007. PLAN: a web platform for automating high-throughput BLAST searches and for managing and mining results. BMC Bioinformatics 8:53.
Hekkelman, M. L., and G. Vriend. 2005. MRS: a fast and compact retrieval system for biological data. Nucleic Acids Res 33:W766-769.
Hemmerich, C., A. Buechlein, R. Podicheti, K. V. Revanna, and Q. Dong. 2010. An Ergatis-based prokaryotic genome annotation web server. Bioinformatics 26:1122-1124.
Henry, C., G. Kerins, J. Blackburn, J. C. Stein, G. C. Smith, D. Eyre, S. Roy, D. Parrott, A. Hart, and S. Goodman. 2009. Defining Environmental Risk Assessment Criteria for Genetically Modified (GM) Mammals and Birds to be placed on the EU market. CT/EFSA/GMO/2009/02.
Herrgard, M. J., M. W. Covert, and B. O. Palsson. 2004. Reconstruction of microbial transcriptional regulatory networks. Curr Opin Biotechnol 15:70-77.
Hillis, D. M., T. A. Heath, and K. St John. 2005. Analysis and visualization of tree space. Syst Biol 54:471-482.
Hittinger, C. T., and S. B. Carroll. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449:677-681.
Holder, M., and P. O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4:275-284.
Holm, L., and P. Rosenstrom. 2010. Dali server: conservation mapping in 3D. Nucleic Acids Res 38:W545-549.
Horton, P., K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, and K. Nakai. 2007. WoLF PSORT: protein localization predictor. Nucleic Acids Res 35:W585-587.
Horvath, P., and R. Barrangou. 2010. CRISPR/Cas, the immune system of bacteria and archaea. Science 327:167-170.
Hult, K., and P. Berglund. 2007. Enzyme promiscuity: mechanism and applications. Trends Biotechnol 25:231-238.
Hunter, S., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res 37:D211-215.
Huson, D. H., D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:460.
Huson, D. H., R. Rupp, and C. Scornavacca. 2011. Phylogenetic Networks: Concepts, Algorithms and Applications Cambridge University Press, Cambridge, UK.
Huwel, S., L. Haalck, N. Conrath, and F. Spener. 1997. Maltose phosphorylase from Lactobacillus brevis: purification, characterization, and application in a biosensor for ortho-phosphate. Enzyme Microb Technol 21:413-420.
Huynen, M., B. Snel, W. Lathe, and P. Bork. 2000. Exploitation of gene context. Curr Opin Struct Biol 10:366-370.
IBCG. 2011. http://www-is.biotoul.fr/. ISfinder. CNRS, LMGM, Universite Louvain la Neuve.
Iliopoulos, I., S. Tsoka, M. A. Andrade, P. Janssen, B. Audit, A. Tramontano, A. Valencia, C. Leroy, C. Sander, and C. A. Ouzounis. 2001. Genome sequences and great expectations. Genome Biol 2:INTERACTIONS0001.
Innis, C. A. 2007. siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. Nucleic Acids Res 35:W489-494.
Inoue, Y., N. Yasutake, Y. Oshima, Y. Yamamoto, T. Tomita, S. Miyoshi, and T. Yatake. 2002. Cloning of the maltose phosphorylase gene from Bacillus sp. strain RK-1 and efficient production of the cloned gene and the trehalose phosphorylase gene from Bacillus stearothermophilus SK-1 in Bacillus subtilis. Biosci Biotechnol Biochem 66:2594-2599.
IUBMB. 1992. Enzyme Nomenclature 1992. Academic Press, San Diego, CA.
Ivics, Z., and Z. Izsvak. 2010. The expanding universe of transposon technologies for gene and cell engineering. Mob DNA 1:25.
Iyer, L. M., L. Aravind, P. Bork, K. Hofmann, A. R. Mushegian, I. B. Zhulin, and E. V. Koonin. 2001. Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. Genome Biol 2:RESEARCH0051.
Jaiswal, P. 2011. Gramene database: a hub for comparative plant genomics. Methods Mol Biol 678:247-275.
Jaroszewski, L., Z. Li, X. H. Cai, C. Weber, and A. Godzik. 2011. FFAS server: novel features and applications. Nucleic Acids Res 39:W38-44.
Jeanmougin, F., J. D. Thompson, M. Gouy, D. G. Higgins, and T. J. Gibson. 1998. Multiple sequence alignment with Clustal X. Trends Biochem Sci 23:403-405.
Johnson, K. L., A. F. Raybould, M. D. Hudson, and G. M. Poppy. 2007. How does scientific risk assessment of GM crops fit within the wider risk analysis?

Trends Plant Sci 12:1-5.

Jones, D. T., and M. B. Swindells. 2002. Getting the most from PSI-BLAST. Trends Biochem Sci 27:161-164.

Jones, M. O., G. D. Koutsovoulos, and M. L. Blaxter. 2011. iPhy: an integrated phylogenetic workbench for supermatrix analyses. BMC Bioinformatics 12:30.

Jones, P., D. Binns, C. McMenamin, C. McAnulla, and S. Hunter. 2011. The InterPro BioMart: federated query and web service access to the InterPro Resource. Database (Oxford) 2011:bar033.

Kaczanowski, S., P. Siedlecki, and P. Zielenkiewicz. 2009. The High Throughput Sequence Annotation Service (HT-SAS) - the shortcut from sequence to true Medline words. BMC Bioinformatics 10:148.

Kaminuma, E., T. Kosuge, Y. Kodama, H. Aono, J. Mashima, T. Gojobori, H. Sugawara, O. Ogasawara, T. Takagi, K. Okubo, and Y. Nakamura. 2011. DDBJ progress report. Nucleic Acids Res 39:D22-27.

Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Res 36:D480-484.

Karp, P. D., S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I. M. Keseler, and R. Caspi. 2010. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform 11:40-79.

Katoh, K., and H. Toh. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286-298.

Kazakov, A. E., M. J. Cipriano, P. S. Novichkov, S. Minovitsky, D. V. Vinogradov, A. Arkin, A. A. Mironov, M. S. Gelfand, and I. Dubchak. 2007. RegTransBase--a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. Nucleic Acids Res 35:D407-412.

Kelder, T., A. R. Pico, K. Hanspers, M. P. van Iersel, C. Evelo, and B. R. Conklin. 2009. Mining biological pathways using WikiPathways web services. PLoS One 4:e6447.

Kelil, A., S. Wang, and R. Brzezinski. 2008. CLUSS2: an alignment-independent algorithm for clustering protein families with multiple biological functions. Int J Comput Biol Drug Des 1:122-140.

Kerkhoven, R., F. H. van Enckevort, J. Boekhorst, D. Molenaar, and R. J. Siezen. 2004. Visualization for genomics: the Microbial Genome Viewer. Bioinformatics 20:1812-1814.

Keseler, I. M., J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muniz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, and P. D. Karp. 2011. EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic Acids Res 39:D583-590.

Khersonsky, O., and D. S. Tawfik. 2010. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem 79:471-505.

Kim, S., E. E. Bolton, and S. H. Bryant. 2011. PubChem3D: Biologically relevant 3-D similarity. J Cheminform 3:26.

Klein, R. J., and S. R. Eddy. 2003. RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics 4:44.

Klimke, W., R. Agarwala, A. Badretdin, S. Chetvernin, S. Ciufo, B. Fedorov, B. Kiryutin, K. O'Neill, W. Resch, S. Resenchuk, S. Schafer, I. Tolstoy, and T. Tatusova. 2009. The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Res 37:D216-223.

Kluyver, A. J. 1924. Eenheid en verscheidenheid in de stofwisseling der microben (Unity and diversity in the metabolism of microorganisms). Chem. Wkbld. 21:266–277.

Kluyver, A. J., and H. J. L. Donker. 1926. Die Einheit in der Biochemie. Chem. Zelle Gewebe 13:134-190.

Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309-338.

Krallinger, M., A. Valencia, and L. Hirschman. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol 9 Suppl 2:S8.

Krieger, E., S. B. Nabuurs, and G. Vriend. 2003. Homology modeling. Methods Biochem Anal 44:509-523.

Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567-580.

Kryshtafovych, A., O. Krysko, P. Daniluk, Z. Dmytriv, and K. Fidelis. 2009. Protein structure prediction center in CASP8. Proteins 77 Suppl 9:5-9.

Kuiper, H. A., and H. V. Davies. 2010. The safe foods risk analysis framework suitable for GMOs? a case study. Food control 21:1662-1676.

Kumar, P., S. Henikoff, and P. C. Ng. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073-1081.

Kuzniar, A., R. C. van Ham, S. Pongor, and J. A. Leunissen. 2008. The quest for orthologs: finding the corresponding gene across genomes. Trends Genet 24:539-551.

Langille, M. G., and F. S. Brinkman. 2009. Bioinformatic detection of horizontally transferred DNA in bacterial genomes. F1000 Biol Rep 1:25.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-2948.

Laskowski, R. A., J. D. Watson, and J. M. Thornton. 2005. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33:W89-93.

Lassmann, T., O. Frings, and E. L. Sonnhammer. 2009. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res 37:858-865.

Lederberg, J. 1948. Gene control of beta-galactosidase in Escherichia coli. Genetics 33:617.

Lees, J., C. Yeats, O. Redfern, A. Clegg, and C. Orengo. 2010. Gene3D: merging structure and function for a

Thousand genomes. Nucleic Acids Res 38:D296-300.

Leinonen, R., R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tarraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, G. Hoad, M. Jang, N. Pakseresht, S. Plaister, R. Radhakrishnan, K. Reddy, S. Sobhany, P. Ten Hoopen, R. Vaughan, V. Zalunin, and G. Cochrane. 2011. The European Nucleotide Archive. Nucleic Acids Res 39:D28-31.

Lemey, P., M. Salemi, and A. Vandamme. 2009. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambridge University Press, Cambridge, UK.

Letunic, I., and P. Bork. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res 39:W475-478.

Letunic, I., and P. Bork. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127-128.

Letunic, I., T. Doerks, and P. Bork. 2009. SMART 6: recent updates and new developments. Nucleic Acids Res 37:D229-232.

Levasseur, A., P. Pontarotti, O. Poch, and J. D. Thompson. 2008. Strategies for reliable exploitation of evolutionary concepts in high throughput biology. Evol Bioinform Online 4:121-137.

Li, Y., and Z. Chen. 2008. RAPD: a database of recombinantly-produced antimicrobial peptides. FEMS Microbiol Lett 289:126-129.

Liang, J. C., R. J. Bloom, and C. D. Smolke. 2011. Engineering biological systems with synthetic RNA molecules. Mol Cell 43:915-926.

Lima, T., A. H. Auchincloss, E. Coudert, G. Keller, K. Michoud, C. Rivoire, V. Bulliard, E. de Castro, C. Lachaize, D. Baratin, I. Phan, L. Bougueleret, and A. Bairoch. 2009. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Res 37:D471-478.

Linard, B., J. D. Thompson, O. Poch, and O. Lecompte. 2011. OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics 12:11.

Liolios, K., I. M. Chen, K. Mavromatis, N. Tavernarakis, P. Hugenholtz, V. M. Markowitz, and N. C. Kyrpides. 2010. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 38:D346-354.

Liu, B., and M. Pop. 2009. ARDB--Antibiotic Resistance Genes Database. Nucleic Acids Res 37:D443-447.

Liu, F., G. Baggerman, L. Schoofs, and G. Wets. 2008. The construction of a bioactive peptide database in Metazoa. J Proteome Res 7:4119-4131.

Liu, G. X., J. Kong, W. W. Lu, W. T. Kong, H. Tian, X. Y. Tian, and G. C. Huo. 2011. beta-Galactosidase with transgalactosylation activity from Lactobacillus fermentum K4. J Dairy Sci 94:5811-5820.

Liu, W., A. Srivastava, and J. Zhang. 2011. A mathematical framework for protein structure comparison. PLoS Comput Biol 7:e1001075.

Liu, X., D. L. Brutlag, and J. S. Liu. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput:127-138.

Lu, Z. 2011. PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford) 2011:baq036.

Ma, H., A. Sorokin, A. Mazein, A. Selkov, E. Selkov, O. Demin, and I. Goryanin. 2007. The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol 3:135.

Ma, S., I. Saaem, and J. Tian. 2011. Error correction in gene synthesis technology. Trends Biotechnol.

Marchler-Bauer, A., S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, and S. H. Bryant. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39:D225-229.

Margelevicius, M., and C. Venclovas. 2010. Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. BMC Bioinformatics 11:89.

Markowitz, V. M., I. M. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang, P. Williams, M. Huntemann, I. Anderson, K. Mavromatis, N. N. Ivanova, and N. C. Kyrpides. 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40:D115-122.

Martinez-Guerrero, C. E., R. Ciria, C. Abreu-Goodger, G. Moreno-Hagelsieb, and E. Merino. 2008. GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. Nucleic Acids Res 36:W176-180.

Maruta, K., K. Mukai, H. Yamashita, M. Kubota, H. Chaen, S. Fukuda, and M. Kurimoto. 2002. Gene encoding a trehalose phosphorylase from Thermoanaerobacter brockii ATCC 35047. Biosci Biotechnol Biochem 66:1976-1980.

Masquida, B., B. Beckert, and F. Jossinet. 2010. Exploring RNA structure by integrative molecular modelling. N Biotechnol 27:170-183.

Matys, V., O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34:D108-110.

Mayr, G., F. S. Domingues, and P. Lackner. 2007. Comparative analysis of protein structure alignments. BMC Struct Biol 7:50.

Mazumder, R., and S. Vasudevan. 2008. Structure-guided comparative analysis of proteins: principles, tools, and applications for predicting function. PLoS Comput Biol 4:e1000151.

McEntyre, J. R., S. Ananiadou, S. Andrews, W. J. Black, R. Boulderstone, P. Buttery, D. Chaplin, S. Chevuru, N. Cobley, L. A. Coleman, P. Davey, B. Gupta, L. Haji-Gholam, C. Hawkins, A. Horne, S. J. Hubbard, J. H.

Kim, I. Lewin, V. Lyte, R. MacIntyre, S. Mansoor, L. Mason, J. McNaught, E. Newbold, C. Nobata, E. Ong, S. Pillai, D. Rebholz-Schuhmann, H. Rosie, R. Rowbotham, C. J. Rupp, P. Stoehr, and P. Vaughan. 2011. UKPMC: a full text article resource for the life sciences. Nucleic Acids Res 39:D58-65.

Meng, Y., C. Shao, H. Wang, and M. Chen. 2011. The regulatory activities of plant microRNAs: a more dynamic perspective. Plant Physiol 157:1583-1595.

Mezard, C. 2006. Meiotic recombination hotspots in plants. Biochem Soc Trans 34:531-534.

Mihara, M., T. Itoh, and T. Izawa. 2010. SALAD database: a motif-based database of protein annotations for plant comparative genomics. Nucleic Acids Res 38:D835-842.

Minkiewicz, P., J. Dziuba, A. Iwaniak, M. Dziuba, and M. Darewicz. 2008. BIOPEP database and other programs for processing bioactive peptide sequences. J AOAC Int 91:965-980.

Mochida, K., and K. Shinozaki. 2011. Advances in omics and bioinformatics tools for systems analyses of plant functions. Plant Cell Physiol 52:2017-2038.

Mochida, K., and K. Shinozaki. 2010. Genomics and bioinformatics resources for crop improvement. Plant Cell Physiol 51:497-523.

Monod, J., and F. Jacob. 1961. General Conclusions: Teleonomic Mechanisms in Cellular Metabolism, Growth and Differentiation. Cold Spring Harbor Symposia On Quantitative Biology 26:389–401.

Moriya, Y., M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182-185.

Morrison, D. A. 1996. Phylogenetic tree-building. Int J Parasitol 26:589-617.

Moult, J., K. Fidelis, A. Kryshtafovych, and A. Tramontano. 2011. Critical assessment of methods of protein structure prediction (CASP)--round IX. Proteins 79 Suppl 10:1-5.

Muller, H. M., E. E. Kenny, and P. W. Sternberg. 2004. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2:e309.

Nain, V., S. Sahi, and P. A. Kumar. 2011. In Silico Identification of Regulatory Elements in Promoters *in* H. S. Lopes, and L. Magalhães Cruz, eds. Computational Biology and Applied Bioinformatics. In Tech, Rijeka, Croatia.

Nakai, H., M. J. Baumann, B. O. Petersen, Y. Westphal, H. Schols, A. Dilokpimol, M. A. Hachem, S. J. Lahtinen, J. O. Duus, and B. Svensson. 2009. The maltodextrin transport system and metabolism in Lactobacillus acidophilus NCFM and production of novel alpha-glucosides through reverse phosphorolysis by maltose phosphorylase. FEBS J 276:7353-7365.

Nakai, H., B. O. Petersen, Y. Westphal, A. Dilokpimol, M. Abou Hachem, J. O. Duus, H. A. Schols, and B. Svensson. 2010. Rational engineering of Lactobacillus acidophilus NCFM maltose phosphorylase into either trehalose or kojibiose dual specificity phosphorylase. Protein Eng Des Sel 23:781-787.

Nath, A., and W. M. Atkins. 2008. A quantitative index of substrate promiscuity. Biochemistry 47:157-166.

NCBI. 2011, posting date. http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html . NCBI. [Online.].

Neph, S., and M. Tompa. 2006. MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes. Nucleic Acids Res 34:W366-368.

Ng, P., and U. Keich. 2008. GIMSAN: a Gibbs motif finder with significance analysis. Bioinformatics 24:2256-2257.

Nguyen, T. H., B. Splechtna, S. Krasteva, W. Kneifel, K. D. Kulbe, C. Divne, and D. Haltrich. 2007a. Characterization and molecular cloning of a heterodimeric beta-galactosidase from the probiotic strain Lactobacillus acidophilus R22. FEMS Microbiol Lett 269:136-144.

Nguyen, T. H., B. Splechtna, M. Yamabhai, D. Haltrich, and C. Peterbauer. 2007b. Cloning and expression of the beta-galactosidase genes from Lactobacillus reuteri in Escherichia coli. J Biotechnol 129:581-591.

Nielsen, J., and M. C. Jewett. 2008. Impact of systems biology on metabolic engineering of Saccharomyces cerevisiae. FEMS Yeast Res 8:122-131.

Nihira, T., H. Nakai, K. Chiku, and M. Kitaoka. 2011. Discovery of nigerose phosphorylase from Clostridium phytofermentans. Appl Microbiol Biotechnol.

Notebaart, R. A., M. A. Huynen, B. Teusink, R. J. Siezen, and B. Snel. 2005. Correlation between sequence conservation and the genomic context after gene duplication. Nucleic Acids Res 33:6164-6171.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205-217.

Novichkov, P. S., O. N. Laikova, E. S. Novichkova, M. S. Gelfand, A. P. Arkin, I. Dubchak, and D. A. Rodionov. 2010. RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. Nucleic Acids Res 38:D111-118.

Oberhardt, M. A., B. O. Palsson, and J. A. Papin. 2009. Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320.

Oberto, J. 2010. FITBAR: a web tool for the robust prediction of prokaryotic regulons. BMC Bioinformatics 11:554.

Okamura, K. 2011. Diversity of animal small RNA pathways and their biological utility. Wiley Interdiscip Rev RNA.

Okumura, T., H. Makiguchi, Y. Makita, R. Yamashita, and K. Nakai. 2007. Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions. Nucleic Acids Res 35:W227-231.

Ooi, H. S., C. Y. Kwo, M. Wildpaner, F. L. Sirota, B. Eisenhaber, S. Maurer-Stroh, W. C. Wong, A. Schleiffer, F. Eisenhaber, and G. Schneider. 2009. ANNIE: integrated de novo protein sequence annotation. Nucleic Acids Res 37:W435-440.

Orth, J. D., T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. O. Palsson. 2011. A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. Mol Syst Biol 7:535.

Osterlund, T., I. Nookaew, and J. Nielsen. 2011. Fifteen years

of large scale metabolic modeling of yeast: Developments and impacts. Biotechnol Adv.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33:5691-5702.

Pagni, M., and C. V. Jongeneel. 2001. Making sense of score statistics for sequence alignments. Brief Bioinform 2:51-67.

Palazzoli, F., F. X. Testu, F. Merly, and Y. Bigot. 2010. Transposon tools: worldwide landscape of intellectual property and technological developments. Genetica 138:285-299.

Pan, Y., C. J. Tsai, B. Ma, and R. Nussinov. 2010. Mechanisms of transcription factor selectivity. Trends Genet 26:75-83.

Paoletti, C., E. Flamm, W. Yan, S. Meek, S. Renckens, M. Fellous, and H. Kuiper. 2008. GMO risk assessment around the world: Some examples. Trends in Food Science & Technology 19:S70–S78.

Park, A. R., and D. K. Oh. 2010. Galacto-oligosaccharide production using microbial beta-galactosidase: current state and perspectives. Appl Microbiol Biotechnol 85:1279-1286.

Passarge, E., B. Horsthemke, and R. A. Farber. 1999. Incorrect use of the term synteny. Nat Genet 23:387.

Pavlopoulou, A., and I. Michalopoulos. 2011. State-of-the-art bioinformatics protein structure prediction tools (Review). Int J Mol Med 28:295-310.

Pearson, W. R. 1995. Comparison of methods for searching protein sequence databases. Protein Sci 4:1145-1160.

Pei, J. 2008. Multiple protein sequence alignment. Curr Opin Struct Biol 18:382-386.

Pei, J., and N. V. Grishin. 2007. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23:802-808.

Pei, J., B. H. Kim, and N. V. Grishin. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36:2295-2300.

Petersen, T. N., S. Brunak, G. von Heijne, and H. Nielsen. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785-786.

Peterson, J. D., L. A. Umayam, T. Dickinson, E. K. Hickey, and O. White. 2001. The Comprehensive Microbial Resource. Nucleic Acids Res 29:123-125.

Petkau, A., M. Stuart-Edwards, P. Stothard, and G. Van Domselaar. 2010. Interactive microbial genome visualization with GView. Bioinformatics 26:3125-3126.

Pico, A. R., T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo. 2008. WikiPathways: pathway editing for the people. PLoS Biol 6:e184.

Pieper, U., B. M. Webb, D. T. Barkan, D. Schneidman-Duhovny, A. Schlessinger, H. Braberg, Z. Yang, E. C. Meng, E. F. Pettersen, C. C. Huang, R. S. Datta, P. Sampathkumar, M. S. Madhusudhan, K. Sjolander, T. E. Ferrin, S. K. Burley, and A. Sali. 2011. ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 39:D465-474.

Pirovano, W., K. A. Feenstra, and J. Heringa. 2008. PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. Bioinformatics 24:492-497.

Pitkanen, E., J. Rousu, and E. Ukkonen. 2010. Computational methods for metabolic reconstruction. Curr Opin Biotechnol 21:70-77.

Planson, A. G., P. Carbonell, E. Paillard, N. Pollet, and J. L. Faulon. 2011. Compound toxicity screening and structure-activity relationship modeling in Escherichia coli. Biotechnol Bioeng 109:846-850.

Podicheti, R., and Q. Dong. 2010. Using WebGBrowse to visualize genome annotation on GBrowse. Cold Spring Harb Protoc 2010:pdb prot5392.

Poolman, B., J. Knol, C. van der Does, P. J. Henderson, W. J. Liang, G. Leblanc, T. Pourcher, and I. Mus-Veteau. 1996. Cation and sugar selectivity determinants in a novel family of transport proteins. Mol Microbiol 19:911-922.

Poolman, B., T. J. Royer, S. E. Mainzer, and B. F. Schmidt. 1989. Lactose transport system of Streptococcus thermophilus: a hybrid protein with homology to the melibiose carrier and enzyme III of phosphoenolpyruvate-dependent phosphotransferase systems. J Bacteriol 171:244-253.

Postma, P. W., J. W. Lengeler, and G. R. Jacobson. 1993. Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. Microbiol Rev 57:543-594.

Powell, S., D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, and P. Bork. 2012. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res 40:D284-289.

Price, N. D., J. A. Papin, C. H. Schilling, and B. O. Palsson. 2003. Genome-scale microbial in silico models: the constraints-based approach. Trends Biotechnol 21:162-169.

Price, N. D., J. L. Reed, and B. O. Palsson. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886-897.

Quester, S., and D. Schomburg. 2011. EnzymeDetector: an integrated enzyme function prediction tool and database. BMC Bioinformatics 12:376.

Rashid, M., D. Singla, A. Sharma, M. Kumar, and G. P. Raghava. 2009. Hmrbase: a database of hormones and their receptors. BMC Genomics 10:307.

Rastogi, S., and B. Rost. 2010. Bioinformatics predictions of localization and targeting. Methods Mol Biol 619:285-305.

Rauch, P. J., M. M. Beerthuyzen, and W. M. de Vos. 1990.

Nucleotide sequence of IS904 from Lactococcus lactis subsp. lactis strain NIZO R5. Nucleic Acids Res 18:4253-4254.

Reading, N. C., D. A. Rasko, A. G. Torres, and V. Sperandio. 2009. The two-component system QseEF and the membrane protein QseG link adrenergic and stress sensing to bacterial pathogenesis. Proc Natl Acad Sci U S A 106:5889-5894.

Reed, J. L., I. Famili, I. Thiele, and B. O. Palsson. 2006. Towards multidimensional genome annotation. Nat Rev Genet 7:130-141.

Reichenbach, B., Y. Gopel, and B. Gorke. 2009. Dual control by perfectly overlapping sigma 54- and sigma 70-promoters adjusts small RNA GlmY expression to different environmental signals. Mol Microbiol 74:1054-1070.

Remm, M., C. E. Storm, and E. L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041-1052.

Rigoutsos, I., T. Huynh, A. Floratos, L. Parida, and D. Platt. 2002. Dictionary-driven protein annotation. Nucleic Acids Res 30:3901-3916.

Rodionov, D. A. 2007. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. Chem Rev 107:3467-3497.

Rogers, D. J., and T. T. Tanimoto. 1960. A Computer Program for Classifying Plants. Science 132:1115-1118.

Romeis, J., D. Bartsch, F. Bigler, M. P. Candolfi, M. M. Gielkens, S. E. Hartley, R. L. Hellmich, J. E. Huesing, P. C. Jepson, R. Layton, H. Quemada, A. Raybould, R. I. Rose, J. Schiemann, M. K. Sears, A. M. Shelton, J. Sweet, Z. Vaituzis, and J. D. Wolt. 2008. Assessment of risk of insect-resistant transgenic crops to nontarget arthropods. Nat Biotechnol 26:203-208.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Rose, P. W., B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, J. D. Westbrook, J. Young, B. Yukich, C. Zardecki, H. M. Berman, and P. E. Bourne. 2011. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39:D392-401.

Rost, B. 2002. Enzyme function less conserved than anticipated. J Mol Biol 318:595-608.

Sadreyev, R. I., M. Tang, B. H. Kim, and N. V. Grishin. 2009. COMPASS server for homology detection: improved statistical accuracy, speed and functionality. Nucleic Acids Res 37:W90-94.

Saier, M. H., Jr., C. V. Tran, and R. D. Barabote. 2006. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res 34:D181-186.

Saier, M. H., Jr., M. R. Yen, K. Noto, D. G. Tamang, and C. Elkan. 2009. The Transporter Classification Database: recent advances. Nucleic Acids Res 37:D274-278.

Salazar, G. A., R. C. Jimenez, A. Garcia, H. Hermjakob, N. Mulder, and E. Blake. 2011. DAS writeback: a collaborative annotation system. BMC Bioinformatics 12:143.

Sanchez, R., F. Serra, J. Tarraga, I. Medina, J. Carbonell, L. Pulido, A. de Maria, S. Capella-Gutierrez, J. Huerta-Cepas, T. Gabaldon, J. Dopazo, and H. Dopazo. 2011. Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. Nucleic Acids Res 39:W470-474.

Satish Kumar, V., M. S. Dasika, and C. D. Maranas. 2007. Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics 8:212.

Scheer, M., A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Sohngen, M. Stelzer, J. Thiele, and D. Schomburg. 2011. BRENDA, the enzyme information system in 2011. Nucleic Acids Res 39:D670-676.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-504.

Schneider, T. D., G. D. Stormo, L. Gold, and A. Ehrenfeucht. 1986. Information content of binding sites on nucleotide sequences. J Mol Biol 188:415-431.

Schnoes, A. M., S. D. Brown, I. Dodevski, and P. C. Babbitt. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5:e1000605.

Schwab, C., K. I. Sorensen, and M. G. Ganzle. 2010. Heterologous expression of glycoside hydrolase family 2 and 42 beta-galactosidases of lactic acid bacteria in Lactococcus lactis. Syst Appl Microbiol 33:300-307.

Scornavacca, C., F. Zickmann, and D. H. Huson. 2011. Tanglegrams for rooted phylogenetic trees and networks. Bioinformatics 27:i248-256.

Seebah, S., A. Suresh, S. Zhuo, Y. H. Choong, H. Chua, D. Chuon, R. Beuerman, and C. Verma. 2007. Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. Nucleic Acids Res 35:D265-268.

Sekizuka, T., M. Matsui, K. Yamane, F. Takeuchi, M. Ohnishi, A. Hishinuma, Y. Arakawa, and M. Kuroda. 2011. Complete sequencing of the bla(NDM-1)-positive IncA/C plasmid from Escherichia coli ST38 isolate suggests a possible origin from plant pathogens. PLoS One 6:e25334.

Servant, F., C. Bru, S. Carrere, E. Courcelle, J. Gouzy, D. Peyruc, and D. Kahn. 2002. ProDom: automated clustering of homologous domains. Brief Bioinform 3:246-251.

Sierro, N., Y. Makita, M. de Hoon, and K. Nakai. 2008. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res 36:D93-96.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539.

Siezen, R. J., and S. A. van Hijum. 2011. Genome (re-

)annotation and open-source annotation pipelines. Microb Biotechnol 3:362-369.

Sigrist, C. J., L. Cerutti, E. de Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, and N. Hulo. 2010. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38:D161-166.

Silvestroni, A., C. Connes, F. Sesma, G. S. De Giori, and J. C. Piard. 2002. Characterization of the melA locus for alpha-galactosidase in Lactobacillus plantarum. Appl Environ Microbiol 68:5464-5471.

Smith, G. R. 1994. Hotspots of homologous recombination. Experientia 50:234-241.

Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. J Mol Biol 147:195-197.

Soding, J., A. Biegert, and A. N. Lupas. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244-248.

Soding, J., M. Remmert, A. Biegert, and A. N. Lupas. 2006. HHsenser: exhaustive transitive profile search using HMM-HMM comparison. Nucleic Acids Res 34:W374-378.

Sparrow, P. A. 2010. GM risk assessment. Mol Biotechnol 44:267-275.

Spok, A., R. M. Twyman, R. Fischer, J. K. Ma, and P. A. Sparrow. 2008. Evolution of a regulatory framework for pharmaceuticals derived from genetically modified plants. Trends Biotechnol 26:506-517.

Staden, R. 1989a. Methods for discovering novel motifs in nucleic acid sequences. Comput Appl Biosci 5:293-298.

Staden, R. 1989b. Methods for calculating the probabilities of finding patterns in sequences. Comput Appl Biosci 5:89-96.

Stano, M., and L. Klucar. 2011. phiGENOME: an integrative navigation throughout bacteriophage genomes. Genomics 98:376-380.

Stoebel, D. M., K. Hokamp, M. S. Last, and C. J. Dorman. 2009. Compensatory evolution of gene regulation in response to stress by Escherichia coli lacking RpoS. PLoS Genet 5:e1000671.

Stormo, G. D. 2000. DNA binding sites: representation and discovery. Bioinformatics 16:16-23.

Storz, G., J. Vogel, and K. M. Wassarman. 2011. Regulation by small RNAs in bacteria: expanding frontiers. Mol Cell 43:880-891.

Suzuki, H., H. Yano, C. J. Brown, and E. M. Top. 2010. Predicting plasmid promiscuity based on genomic signature. J Bacteriol 192:6045-6055.

Szklarczyk, D., A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39:D561-568.

Tamis, W. L. M., A. van Dommelen, and G. R. de Snoo. 2009. Lack of transparency on environmental risks of genetically modified micro-organisms in industrial biotechnology. Journal of cleaner production 17:581-592.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739.

Tatusov, R. L., S. F. Altschul, and E. V. Koonin. 1994. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci U S A 91:12091-12095.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

Tengs, T., A. B. Kristoffersen, K. G. Berdal, T. Thorstensen, M. A. Butenko, H. Nesvold, and A. Holst-Jensen. 2007. Microarray-based method for detection of unknown genetic modifications. BMC Biotechnol 7:91.

Tengs, T., H. Zhang, A. Holst-Jensen, J. Bohlin, M. A. Butenko, A. B. Kristoffersen, H. G. Sorteberg, and K. G. Berdal. 2009. Characterization of unknown genetic modifications using high throughput sequencing and computational subtraction. BMC Biotechnol 9:87.

Teusink, B., H. V. Westerhoff, and F. J. Bruggeman. 2010. Comparative systems biology: from bacteria to man. Wiley Interdiscip Rev Syst Biol Med 2:518-532.

The Uniprot Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 40:D71-D75

Thomas, P. D., M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129-2141.

Thompson, J. D., B. Linard, O. Lecompte, and O. Poch. 2011. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PLoS One 6:e18093.

Thompson, W., E. C. Rouchka, and C. E. Lawrence. 2003. Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res 31:3580-3585.

Trachana, K., T. A. Larsson, S. Powell, W. H. Chen, T. Doerks, J. Muller, and P. Bork. 2011. Orthology prediction methods: a quality assessment using curated protein families. Bioessays 33:769-780.

Tucker, B. J., and R. R. Breaker. 2005. Riboswitches as versatile gene control elements. Curr Opin Struct Biol 15:342-348.

Tudor, C. O., C. J. Schmidt, and K. Vijay-Shanker. 2010. eGIFT: mining gene information from the literature. BMC Bioinformatics 11:418.

US government. 2010. Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA *in* U. government, ed, http://www.phe.gov/Preparedness/legal/guidance/syndna/Documents/syndna-guidance.pdf.

Vallenet, D., S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli, and C. Medigue. 2009. MicroScope: a platform for microbial genome annotation and comparative genomics. Database (Oxford) 2009:bap021.

van der Heijden, R. T., B. Snel, V. van Noort, and M. A. Huynen. 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics 8:83.

Van Domselaar, G. H., P. Stothard, S. Shrivastava, J. A. Cruz, A. Guo, X. Dong, P. Lu, D. Szafron, R. Greiner, and D. S. Wishart. 2005. BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res 33:W455-459.

Van Hellemont, R., P. Monsieurs, G. Thijs, B. de Moor, Y. Van de Peer, and K. Marchal. 2005. A novel approach to identifying regulatory motifs in distantly related genomes. Genome Biol 6:R113.

Varian, H. 2005. Bootstrap Tutorial. Mathematica J. 9:768-775.

Varshney, R. K., S. N. Nayak, G. D. May, and S. A. Jackson. 2009. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27:522-530.

Venselaar, H., R. P. Joosten, B. Vroling, C. A. Baakman, M. L. Hekkelman, E. Krieger, and G. Vriend. 2010a. Homology modelling and spectroscopy, a never-ending love story. Eur Biophys J 39:551-563.

Venselaar, H., T. A. Te Beek, R. K. Kuipers, M. L. Hekkelman, and G. Vriend. 2010b. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics 11:548.

Vitreschak, A. G., A. A. Mironov, V. A. Lyubetsky, and M. S. Gelfand. 2008. Comparative genomic analysis of T-box regulatory systems in bacteria. RNA 14:717-735.

Vitreschak, A. G., D. A. Rodionov, A. A. Mironov, and M. S. Gelfand. 2004. Riboswitches: the oldest mechanism for the regulation of gene expression? Trends Genet 20:44-50.

Vlahovicek, K., L. Kajan, J. Murvai, Z. Hegedus, and S. Pongor. 2003. The SBASE domain sequence library, release 10: domain architecture prediction. Nucleic Acids Res 31:403-405.

von Grotthuss, M., D. Plewczynski, G. Vriend, and L. Rychlewski. 2008. 3D-Fun: predicting enzyme function from structure. Nucleic Acids Res 36:W303-307.

Walsh, T. R., J. Weeks, D. M. Livermore, and M. A. Toleman. 2011. Dissemination of NDM-1 positive bacteria in the New Delhi environment and its implications for human health: an environmental point prevalence study. Lancet Infect Dis 11:355-362.

Wan, Y., M. Kertesz, R. C. Spitale, E. Segal, and H. Y. Chang. 2011. Understanding the transcriptome through RNA structure. Nat Rev Genet 12:641-655.

Wang, G., X. Li, and Z. Wang. 2009. APD2: the updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res 37:D933-937.

Wassenaar, T. M., and W. Gaastra. 2001. Bacterial virulence: can we draw the line? FEMS Microbiol Lett 201:1-7.

Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189-1191.

Watson, J. D., R. A. Laskowski, and J. M. Thornton. 2005. Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15:275-284.

Weber, W., and M. Fussenegger. 2011. Emerging biomedical applications of synthetic biology. Nat Rev Genet 13:21-35.

Wels, M., C. Francke, R. Kerkhoven, M. Kleerebezem, and R. J. Siezen. 2006. Predicting cis-acting elements of Lactobacillus plantarum by comparative genomics with different taxonomic subgroups. Nucleic Acids Res 34:1947-1958.

Wels, M., T. Groot Kormelink, M. Kleerebezem, R. J. Siezen, and C. Francke. 2008. An in silico analysis of T-box regulated genes and T-box evolution in prokaryotes, with emphasis on prediction of substrate specificity of transporters. BMC Genomics 9:330.

Westhof, E., B. Masquida, and F. Jossinet. 2011. Predicting and modeling RNA architecture. Cold Spring Harb Perspect Biol 3.

Whelan, S. 2008. Inferring trees. Methods Mol Biol 452:287-309.

Whisstock, J. C., and A. M. Lesk. 2003. Prediction of protein function from protein sequence and structure. Q Rev Biophys 36:307-340.

Wilson, D., V. Charoensawan, S. K. Kummerfeld, and S. A. Teichmann. 2008. DBD--taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res 36:D88-92.

Wilson, D., R. Pethica, Y. Zhou, C. Talbot, C. Vogel, M. Madera, C. Chothia, and J. Gough. 2009. SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res 37:D380-386.

Wolf, M. Y., Y. I. Wolf, and E. V. Koonin. 2008. Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. Biol Direct 3:40.

Wolfsberg, T. G. 2011. Using the NCBI Map Viewer to browse genomic sequence data. Curr Protoc Hum Genet Chapter 18:Unit18 15.

Wolfsberg, T. G. 2010. Using the NCBI map viewer to browse genomic sequence data. Curr Protoc Bioinformatics Chapter 1:Unit 1 5 1-25.

Wong, K. M., M. A. Suchard, and J. P. Huelsenbeck. 2008. Alignment uncertainty and genomic analysis. Science 319:473-476.

Wong, W. C., S. Maurer-Stroh, and F. Eisenhaber. 2010. More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. PLoS Comput Biol 6:e1000867.

Worth, C. L., R. Preissner, and T. L. Blundell. 2011. SDM--a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res 39:W215-222.

Xiao, Y., S. R. Wigneshweraraj, R. Weinzierl, Y. P. Wang, and M. Buck. 2009. Construction and functional analyses of a comprehensive sigma54 site-directed

mutant library using alanine-cysteine mutagenesis. Nucleic Acids Res 37:4482-4497.

Yakunin, A. F., A. A. Yee, A. Savchenko, A. M. Edwards, and C. H. Arrowsmith. 2004. Structural proteomics: a tool for genome annotation. Curr Opin Chem Biol 8:42-48.

Yamamoto, T., K. Maruta, K. Mukai, H. Yamashita, T. Nishimoto, M. Kubota, S. Fukuda, M. Kurimoto, and Y. Tsujisaka. 2004. Cloning and sequencing of kojibiose phosphorylase gene from Thermoanaerobacter brockii ATCC35047. J Biosci Bioeng 98:99-106.

Yang, J., L. Chen, L. Sun, J. Yu, and Q. Jin. 2008. VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Res 36:D539-542.

Yi, T. M., and E. S. Lander. 1994. Recognition of related proteins by iterative template refinement (ITR). Protein Sci 3:1315-1328.

Yong, D., M. A. Toleman, C. G. Giske, H. S. Cho, K. Sundman, K. Lee, and T. R. Walsh. 2009. Characterization of a new metallo-beta-lactamase gene, bla(NDM-1), and a novel erythromycin esterase gene carried on a unique genetic structure in Klebsiella pneumoniae sequence type 14 from India. Antimicrob Agents Chemother 53:5046-5054.

Youens-Clark, K., E. Buckler, T. Casstevens, C. Chen, G. Declerck, P. Derwent, P. Dharmawardhana, P. Jaiswal, P. Kersey, A. S. Karthikeyan, J. Lu, S. R. McCouch, L. Ren, W. Spooner, J. C. Stein, J. Thomason, S. Wei, and D. Ware. 2011. Gramene database in 2010: updates and extensions. Nucleic Acids Res 39:D1085-1094.

Yu, N. Y., J. R. Wagner, M. R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S. C. Sahinalp, M. Ester, L. J. Foster, and F. S. Brinkman. 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26:1608-1615.

Zamyatnin, A. A., A. S. Borchikov, M. G. Vladimirov, and O. L. Voronina. 2006. The EROP-Moscow oligopeptide database. Nucleic Acids Res 34:D261-266.

Zaremba, S., M. Ramos-Santacruz, T. Hampton, P. Shetty, J. Fedorko, J. Whitmore, J. M. Greene, N. T. Perna, J. D. Glasner, G. Plunkett, 3rd, M. Shaker, and D. Pot. 2009. Text-mining of PubMed abstracts by natural language processing to create a public knowledge base on molecular mechanisms of bacterial enteropathogens. BMC Bioinformatics 10:177.

Zhang, N., N. Joly, P. C. Burrows, M. Jovanovic, S. R. Wigneshweraraj, and M. Buck. 2009. The role of the conserved phenylalanine in the sigma54-interacting GAFTGA motif of bacterial enhancer binding proteins. Nucleic Acids Res 37:5981-5992.

Zhang, P., K. Dreher, A. Karthikeyan, A. Chi, A. Pujar, R. Caspi, P. Karp, V. Kirkup, M. Latendresse, C. Lee, L. A. Mueller, R. Muller, and S. Y. Rhee. 2010. Creation of a genome-wide metabolic pathway database for Populus trichocarpa using a new approach for reconstruction and curation of metabolic pathways for plants. Plant Physiol 153:1479-1491.

Zhang, S., W. Su, and J. Yang. 2009. ARCS-Motif: discovering correlated motifs from unaligned biological sequences. Bioinformatics 25:183-189.

Zhou, C. E., J. Smith, M. Lam, A. Zemla, M. D. Dyer, and T. Slezak. 2007. MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res 35:D391-394.

Zhou, H., and Y. Zhou. 2005. SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics 21:3615-3621.

Zhou, M., J. Boekhorst, C. Francke, and R. J. Siezen. 2008. LocateP: genome-scale subcellular-location predictor for bacterial proteins. BMC Bioinformatics 9:173.

Zhu, J., and M. Q. Zhang. 1999. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics 15:607-611.

Zubieta, C., K. A. Arkus, R. E. Cahoon, and J. M. Jez. 2008. A single amino acid change is responsible for evolution of acyltransferase specificity in bacterial methionine biosynthesis. J Biol Chem 283:7561-7567.

TABLE 1. The main public resources of sequence information

name	website	institute	reference	content
genome and protein sequence repositories				
NCBI	www.ncbi.nlm.nih.gov	USA	(Benson et al. 2012)	central repository of genome sequence information
EBI-EMBL	www.ebi.ac.uk	Europe	(Leinonen et al. 2011)	central repository of genome sequence information
DDBJ	www.ddbj.nig.ac.jp	Japan	(Kaminuma et al. 2011)	central repository of genome sequence information
protein databases				
Uniprot knowledgebase	www.ebi.ac.uk/uniprot/index.html	EBI-SIB	(the Uniprot Consortium 2012)	central repository of protein sequence information
Uniprot/SwissProt	web.expasy.org/docs/swiss-prot_guideline.html	EBI-SIB		curated repository of protein sequence information
PDB	www.pdb.org/pdb/home/home.do	RCSB (Rutgers, UCSD and Wisconsin Univ.)	(Rose et al. 2011)	curated repository of protein structures

Remark: A perfect example of the care one has to take when searching public space using sequence identifiers are the gene identifiers attributed by NCBI. Upon data deposit, NCBI attributes unique identifiers to the submitted data and the submitted data are stored as a GenBank file. Then NCBI creates a copy datafile with new unique identifiers called the RefSeq file. All relevant information within the NCBI databases is linked to this RefSeq file and similarly the meta-data in other databases often only refers to this RefSeq gene identifier, resulting in incomplete recovery of data in case the GenBank gene identifier is used.

TABLE 2. Resources of functional classification and classified sequences

name	website	institute	reference	content
classification based on function				
COG	www.ncbi.nlm.nih.gov/COG	NCBI	(Tatusov et al. 2003)	annotated clusters of orthologous sequences
GO	www.geneontology.org	GO consortium	(Ashburner et al. 2000)	function description using gene ontologies
AmiGO			(Carbon et al. 2009)	online access to ontology and annotation data
EC-number	enzyme.expasy.org	SIB	(Bairoch 2000)	enzyme database defining EC-numbers
	www.chem.qmul.ac.uk/iubmb/enzyme	IUBMB	(IUBMB 1992)	EC-number definitions (enzymes)
Cazy	www.cazy.org	AFMB (Marseille)	(Cantarel et al. 2009)	classification system for Carbohydrate-Active EnZymes
TCDB	www.tcdb.org	UCSD	(Saier et al. 2009)	TC-number definitions (transport systems)
classification based on sequence profiles or structure				
CDD	www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	NCBI	(Marchler-Bauer et al. 2011)	conserved protein domain database at NCBI
HAMAP	hamap.expasy.org/	HAMAP	(Lima et al. 2009)	microbial proteome
Interpro	www.ebi.ac.uk/interpro	EBI	(Hunter et al. 2009)	protein families database
PFAM	pfam.sanger.ac.uk	Wellcome Trust Sanger Inst. and Howard Hughes Janelia Farm Res. Campus	(Finn et al. 2010)	protein families database
Prodom	prodom.prabi.fr	INRA/CNRS	(Servant et al. 2002)	protein families database based on Uniprot
ProtClustDB	www.ncbi.nlm.nih.gov/sites/entrez?db=protein clusters	NCBI	(Klimke et al. 2009)	Protein Clusters Database
TIGR fams	cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi	JGI	(Peterson et al. 2001)	protein families database

protein structure				
CATH	www.cathdb.info	Univ. College London	(Greene et al. 2007)	classification of protein domain structures
SCOP	scop.mrc-lmb.cam.ac.uk/scop/	MRC Cambridge	(Andreeva et al. 2008)	Structural Classification of Proteins
transcription factor families				
DBD	dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?Home	MRC Cambridge	(Wilson et al. 2008)	Transcription factor database
functional RNA				
RFAM	rfam.sanger.ac.uk	Wellcome Trust Sanger Inst. and Howard Hughes Janelia Farm Res. Campus	(Gardner et al. 2009)	RNA families database

TABLE 3. Public resources of function information

name	website	institute	reference	content	
genomes					
GOLD	www.genomesonline.org	DOE-JGI and LBNL	(Liolios et al. 2010)	status of genomic and metagenomic projects and their associated metadata	curated, refs
enzymes					
BRENDA	www.brenda-enzymes.org	TU Braunschweig	(Scheer et al. 2011)	comprehensive enzyme information system	cur, refs
Cazy	www.cazy.org	AFMB (Marseille)	(Cantarel et al. 2009)	database of Carbohydrate-Active EnZymes	cur, refs
transporters					
TCDB	www.tcdb.org	UCSD	(Saier et al. 2009)	Functional and Phylogenetic Classification of Membrane Transport Proteins	cur, refs
ABCISSE	www1.pasteur.fr/recherche/unites/pmtg/abc/database.iphtml	Institut Pasteur	(Bouige et al. 2002)	database of ATP-binding cassette (ABC) systems	aut-cur,refs
pathways linked to reactions linked to compounds and proteins/genes					
KEGG	www.genome.jp/kegg	Kyoto Univ.	(Kanehisa et al. 2008)	most comprehensive pathway/genome database for all species	aut-cur
Biocyc	biocyc.org/	SRI Int.; Menlo Park CA	(Caspi et al. 2009)	MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, which includes Ecocyc	aut-cur,refs
Ecocyc	biocyc.org/ecocyc/index.shtml		(Keseler et al. 2011)		cur,refs
PMN	plantcyc.org	Carnegie Inst. Sci.	(Zhang et al. 2010)	general database of metabolic pathways for plants (Pathway Tools system)	aut-cur,refs
Gramene	www.gramene.org/pathway	CSHL, Cornell Univ., Oregon state, USDA	(Jaiswal 2011; Youens-Clark et al. 2011)	The Gramene database is a hub for comparative plant genomics	aut-cur,refs
the SEED	www.theseed.org	Argonne Natl. Lab. and the Univ. of Chicago	(Overbeek et al. 2005)	genome annotations for archaea, bacteria and viruses	aut-cur,refs

pathways					
PathwayCommons	www.pathwaycommons.org/pc	Univ. of Toronto	(Cerami et al. 2011)	a web resource for biological pathway data in animals and yeast	aut-cur
Reactome	www.reactome.org/ReactomeGWT/entrypoint.html	NIH and EU	(D'Eustachio 2011)	knowledgebase of human biological pathways and processes	cur,refs
Wikipathways	wikipathways.org/index.php/WikiPathways	Univ. Maastricht, UCSF	(Pico et al. 2008; Kelder et al. 2009)	a web resource for biological pathway data mainly in animals and yeast	cur,refs

TABLE 4. Automated literature searches

name	website	institute	reference	content
general				
Google scholar	scholar.google.nl	Google	©2011 Google	Google Scholar provides a simple way to broadly search for scholarly literature
PubMed, PMC	www.ncbi.nlm.nih.gov/pubmed	NCBI	(Lu 2011)	resource that searches all life science literature. rem: lacks relevant literature before 2000.
UKPMC	ukpmc.ac.uk	EBI-EMBL, WT, U.Man.	(McEntyre et al. 2011)	a full text article resource for the life sciences. Ext. PMC
Web Of Science	apps.webofknowledge.com	Thomson Reuters	© 2011 Thomson Reuters	cited reference search forward and backward in time
directed				
Bio Dictionary	cbcsrv.watson.ibm.com/Tpa.html	IBM	(Rigoutsos et al. 2002)	Dictionary-driven protein annotation using Watson
BioLit	biolit.ucsd.edu	UCSD	(Fink et al. 2008)	integrating biological literature with databases
eGIFT	biotm.cis.udel.edu/eGIFT	Univ. of Delaware	(Tudor, Schmidt, and Vijay-Shanker 2010)	Mining human Gene Information from the Literature. No retrieval of information on bacterial or plant systems.
e-LiSe	miron.ibb.waw.pl/e-LiSe	IBB, Polish Acad. Sci.	(Gladki et al. 2008)	Service for searching words in publication abstracts.
Textpresso	www.textpresso.org	CalTech, Passadena	(Muller, Kenny, and Sternberg 2004)	an ontology-based information retrieval and extraction system for biological literature
sequence based				
htsas	miron.ibb.waw.pl/htsas	IBB, Polish Acad. of Sci. (Warszawa)	(Kaczanowski, Siedlecki, and Zielenkiewicz 2009)	Service for automatic annotation of proteins using both the sequence and words derived from publication abstracts
PubDNA Finder	servet.dia.fi.upm.es:8080/pubdnafinder	Univ. Politécnica de Madrid	(Garcia-Remesal et al. 2010)	a web database linking full-text articles to sequences of nucleic acids
pubmed2ensembl	www.pubmed2ensembl.org	Univ. of Manchester	(Baran et al. 2011)	search literature for nucleotide sequences specifically for yeast, fungi and animals
text2genome	text2genome.smith.man.ac.uk	Univ. of Manchester	(Haeussler, Gerner, and Bergman 2011)	search literature for nucleotide sequences

TABLE 5. Public resources of virulence factors and antibiotic resistance genes

name	website	institute	reference	content
virulence factors				
VFDB	www.mgc.ac.cn/VFs	China	(Yang et al. 2008)	virulence factors of pathogenic bacteria
MvirDB	mvirdb.llnl.gov	Lawrence Livermore Natl. Lab. USA	(Zhou et al. 2007)	database of microbial protein toxins, virulence factors and antibiotic resistance genes
PHI database	www.phibase.org/	BBSRC (UK)	(Baldwin et al. 2006)	database of proteins involved in Pathogen-Host Interactions
antibiotic resistance genes				
ARDB	ardb.cbcb.umd.edu	Univ. of Maryland USA	(Liu and Pop 2009)	Antibiotic Resistance Genes Database
Arpcard	arpcard.mcmaster.ca	McMaster Univ. Hamilton, Ca	(CARD 2011)	Antibiotic Resistance Database
EcARS	www.broadinstitute.org/annotation/genome/escherichia_antibiotic_resistance/MultiHome.html	Broad Inst.	(EcARS 2011)	*Ecoli* specific Antibiotic Resistance
general information				
EARS-Net	www.ecdc.europa.eu/en/activities/surveillance/EARS-Net	RIVM	(Gagliotti et al. 2011)	EARS-Net is a European wide network of national surveillance systems, providing European reference data on antimicrobial resistance for public health purposes
NMPDR	www.nmpdr.org/FIG/wiki/view.cgi	BRC		The NMPDR provided curated annotations in an environment for comparative analysis of genomes and biological subsystems, with an emphasis on the food-borne pathogens
PFGS	pfgrc.jcvi.org pathogen database	J. Graig Venter Inst.		Pathogen Functional genomics resource
ERIC	www.ericbrc.org	ERIC-BRC, SRA Int. Inc	(Zaremba et al. 2009)	a public knowledge base on molecular mechanisms of bacterial enteropathogens created via Text-mining

TABLE 6. Public resources of bioactive peptides

name	website	reference
bioactive peptides		
AMPer	marray.cmdr.ubc.ca/cgi-bin/amp.pl	
ANTIMIC	research.i2r.a-star.edu.sg/Templar/DB/ANTIMIC	(Brahmachary et al. 2004)
APD	aps.unmc.edu/AP	(Wang, Li, and Wang 2009)
BACTIBASE	bactibase.pfba-lab-tun.org	(Hammami et al. 2010)
BAGEL2	bagel2.molgenrug.nl	(de Jong et al. 2010)
BioPD	biopd.bjmu.edu.cn	(Minkiewicz et al. 2008)
Defensins knowledgebase	defensins.bii.a-star.edu.sg	(Seebah et al. 2007)
EROP-Moscow	erop.inbi.ras.ru	(Zamyatnin et al. 2006)
Hmrbase	crdd.osdd.net/raghava/hmrbase	(Rashid et al. 2009)
NORINE	bioinfo.lifl.fr/norine	(Caboche et al. 2008)
PepBank	pepbank.mgh.harvard.edu	(Duchrow et al. 2009)
PeptideDB	www.peptides.be	(Liu et al. 2008)
PhytAMP	phytamp.pfba-lab-tun.org	(Hammami et al. 2009)
RAPD	faculty.ist.unomaha.edu/chen/rapd	(Li and Chen 2008)
SwePep	www.swepep.org	(Falth et al. 2006)
AMSDb	www.bbcm.units.it/~tossi/pag2.htm	
PenBase	www.penbase.immunaqua.com	(Gueguen et al. 2006)

TABLE 7. Search similar sequences

Tool	Website	Institute	Reference	Purpose
BLAST	blast.ncbi.nlm.nih.gov/Blast.cgi	NCBI refseq	(Altschul et al. 1997)	nucleotide-nucleotide: BLASTN protein-protein: BLASTP or PSI-BLAST protein-nucleotide: TBLASTN
FASTA/SSEARCH/RSEARCH	www.ebi.ac.uk/Tools/sss/fasta or fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml	Uniprot	(Pearson 1995) (Klein and Eddy 2003)	protein RNA
ENA	www.ebi.ac.uk/ena	European Nucleotide Archive	(Leinonen et al. 2011)	Nucleotide Sequence
BOV	bioportal.cgb.indiana.edu/cgi-bin/BOV	Indiana Univ.	(Gollapudi et al. 2008)	a web-based BLAST output visualization tool
MRS	mrs.cmbi.ru.nl/mrs-web	CMBI, RUNijmegen MC	(Hekkelman and Vriend 2005)	A fast and compact retrieval system for biological data
PLAN	bioinfo.noble.org/plan	Samuel Roberts Noble Foundation, Ardmore	(He, Dai, and Zhao 2007)	automating high-throughput BLAST searches and for managing and mining results

TABLE 8. Search for similar sequence profiles

Tool	Website	nstitute	Reference	Content
protein sequence classification				
Superfamily	supfam.org/SUPERFAMILY/index.html	Univ. of Bristol	(Wilson et al. 2009)	
CDD	www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	NCBI	(Marchler-Bauer et al. 2011)	conserved protein domain database at NCBI
COMA	bioinformatics.ibt.lt:8085/coma	Vilnius Univ.	(Margelevicius and Venclovas 2010)	
COMPASS	prodata.swmed.edu/compass	Howard Hughes Med. Inst., Univ. Texas	(Sadreyev et al. 2009)	
DAS server	mendel.imp.ac.at/sat/DAS/DAS.html	Univ. of Birmingham	(Cserzo et al. 2002)	
FFAS	ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl	Sanford|Burham Med.Res. Inst.	(Jaroszewski et al. 2011)	
Gene3D	gene3d.biochem.ucl.ac.uk	UC London	(Lees et al. 2010)	CATH
HHpred/HHsenser	toolkit.tuebingen.mpg.de	MPG, Tuebingen	(Soding et al. 2006)	HMM vs HMM
Interpro	www.ebi.ac.uk/interpro	EBI	(Hunter et al. 2009)	
Panther	www.pantherdb.org	UCSC	(Thomas et al. 2003)	GO related
PFAM	pfam.sanger.ac.uk	WT-Sanger Inst.	(Finn et al. 2010)	
PRIAM	priam.prabi.fr	INRA	(Claudel-Renard et al. 2003)	enzyme detection
Prints	www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php	Univ. of Manchester	(Gaudet et al. 2012)	
Prodom	prodom.prabi.fr/prodom/current/html/home.php	INRA/CNRS, Uniprot	(Servant et al. 2002)	
proFAT	cluster-1.mpi-cbg.de/profat/profat.html	MPI, Dresden	(Bradshaw, Surendranath, and Habermann 2006)	a web-based tool for the functional annotation of protein sequences

ProSite	prosite.expasy.org	SIB/Expasy	(Sigrist et al. 2010)	
SALAD	salad.dna.affrc.go.jp/salad/en	Natl. Inst. of Agrobiological Sci., Tsukuba	(Mihara, Itoh, and Izawa 2010)	
SBASE	hydra.icgeb.trieste.it/sbase	ICGEB, Trieste	(Vlahovicek et al. 2003)	support vector machine based prediction of protein domains
SMART	smart.embl-heidelberg.de	EMBL	(Letunic, Doerks, and Bork 2009)	
protein structure comparison				
Dali	ekhidna.biocenter.helsinki.fi/dali_server	Univ. of Helsinki	(Holm and Rosenstrom 2010)	The Dali server is a network service for comparing protein structures in 3D
HOPE	www.cmbi.ru.nl/hope	CMBI, RUNijmegen MC	(Venselaar et al. 2010b)	prediction of effect mutations through homology protein structure models
MODbase	salilab.org/modbase	UC San Fransisco	(Pieper et al. 2011)	a database of annotated comparative protein structure models and associated resources
PESDserv	reccr.chem.rpi.edu/Software/pesdserv	Rensselaer Polytechnic Inst. Troy	(Das, Krein, and Breneman 2010)	a server for high-throughput comparison of protein binding site surfaces
psipred	bioinf.cs.ucl.ac.uk/psipred	UCL, Bloomsbury	(Bryson et al. 2005)	protein structure predictor
SiteFinder|3D	Sage.csb.yale.edu/sitefinder3d	Howard Hughes Med. Inst., Yale Univ.	(Innis 2007)	Identification of functional sites in a protein structure
protein location				
LocateP	www.cmbi.ru.nl/locatep-db/cgi-bin/locatepdb.py	CMBI, RUNijmegen MC	(Zhou et al. 2008)	Subcellular location prediction for proteins of Gram-positive species
PSORTb v.3.0	www.psort.org/psortb	Simon Fraser Univ.	(Yu et al. 2010)	Prediction of protein location in bacteria and archaea
signalP	www.cbs.dtu.dk/services/SignalP	Techn.l Univ. of Denmark	(Petersen et al. 2011)	signal peptides

TCDB	www.tcdb.org	UC San Diego	(Saier, Tran, and Barabote 2006)	Transporter database and classification
TMHMM	www.cbs.dtu.dk/services	Techn. Univ.of Denmark	(Krogh et al. 2001)	Transmembrane helices
wolf PSORT	wolfpsort.org	CBRC, Japan	(Horton et al. 2007)	Subcellular location of eukaryotic proteins
Annotation of individual proteins				
ANNIE	Annie.bii.a-star.edu.sg	A*Star, Singapore	(Ooi et al. 2009)	Annotation of individual protein on basis of various algorithms
EnzymeDetector	Enzymedetector.tu-bs.de	Techn. Univ. Braunschweig	(Quester and Schomburg 2011)	Annotation of proteins on basis of BRENDA
3D-Fun	3dfun.bioinfo.pl	Bioinfobank Inst., Poznan	(von Grotthuss et al. 2008)	Annotation of proteins on basis of 3D structure
ProFunc	www.ebi.ac.uk/thornton-srv/databases/ProFunc	EBI-WT, Hinxton	(Laskowski, Watson, and Thornton 2005)	Annotation of proteins on basis of 3D structure
SIFT	Sift.jcvi.org	JGVI, San Diego	(Kumar, Henikoff, and Ng 2009)	Identification of residues important for function

TABLE 9. Sequence alignment tools

Tool	Reference	Website
ClustalX/W and Omega	(Jeanmougin et al. 1998; Larkin et al. 2007)	www.clustal.org (software) www.ebi.ac.uk/Tools/msa/clustalo
Kalign	(Lassmann, Frings, and Sonnhammer 2009)	msa.cgb.ki.se (software and server)
T-Coffee/ Expresso	(Notredame, Higgins, and Heringa 2000; Di Tommaso et al. 2011)	http://tcoffee.crg.cat/
MAFFT	(Katoh and Toh 2008)	mafft.cbrc.jp/alignment/software/ (software)
MUSCLE	(Edgar 2004)	www.drive5.com/muscle/ (software)
ProbCons	(Do et al. 2005)	probcons.stanford.edu/
PRALINE	(Pirovano, Feenstra, and Heringa 2008)	www.ibi.vu.nl/programs/pralinewww/
PROMALS(3D)	(Pei, Kim, and Grishin 2008)	prodata.swmed.edu/promals(sD)/promals(3D).php
SPEM(3D)	(Zhou and Zhou 2005)	sparks.informatics.iupui.edu/Softwares-Services_files/spem_3d.htm

TABLE 10. Sequence clustering on basis of MSA

Tool	Website	Institute	Reference	Content
Tree view and analysis				
Dendroscope	www.dendroscope.org	Univ. Tuebingen	(Scornavacca, Zickmann, and Huson 2011)	An interactive viewer for large phylogenetic trees with several new algorithms to re-cluster sequences
iTOL	itol.embl.de	EMBL	(Letunic and Bork 2011)	an online tool for phylogenetic tree display and annotation
LOFT	www.cmbi.ru.nl/LOFT	CMBI, RU Nijmegen MC	(van der Heijden et al. 2007)	Orth. pred. at scalable resolution by phylogenetic tree
generate clusters of sequences				
BLAST-EXPLORER	www.phylogeny.fr	CNRS,Marseille	(Dereeper et al. 2010)	building datasets for phylogenetic analysis
MrBayes	"	"	(Ronquist and Huelsenbeck 2003)	Bayesian analysis of trees
CLUSS	prospectus.usherbrooke.ca/CLUSS	Univ. de Sherbrooke, Canada	(Kelil, Wang, and Brzezinski 2008)	sequence clustering tool
iPhy	iphy.bio.ed.ac.uk	Univ. of Edinburgh	(Jones, Koutsovoulos, and Blaxter 2011)	an integrated phylogenetic workbench for supermatrix analyses
Jalview	www.jalview.org/	BBSRC, Univ. of Dundee	(Waterhouse et al. 2009)	a multiple sequence alignment editor and analysis workbench
MEGA	www.megasoftware.net/	The Biodesign Inst., Tempe	(Tamura et al. 2011)	Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods
Phylemon 2	phylemon.bioinfo.cipf.es/	Centro de Investigación Príncipe Felipe, Valencia	(Sanchez et al. 2011)	a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing
PHYLIP	cmgm.stanford.edu/phylip/	Stanford Univ.	(Felsenstein 2004)	Software package for Phylogeny Inference
PhyML	www.atgc-montpellier.fr/phyml	Univ. Montpellier	(Guindon et al. 2010)	A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood
TREE-PUZZLE	www.tree-puzzle.de	CIBIF,MFPL, Wien	(Schmidt et al. 2002)	maximum likelihood phylogenetic analysis using quartets and parallel computing

TABLE 11. Comparative Genome viewers

Tool	Website	Institute	Reference	Content
Artemis	www.sanger.ac.uk/resources/software/artemis/	Sanger Inst.	(Carver et al. 2008)	An interactive downloadable viewer
autoGRAPH	genoweb.univ-rennes1.fr /tom_dog/AutoGRAPH	Univ. Rennes	(Derrien et al. 2007)	an interactive web server for automating and visualizing comparative genome maps
Ensembl browser	www.ensembl.org	EMBL-EBI	(Fernandez-Suarez and Schuster 2010)	server to browse genomic sequence data
GeConT II	bioinfo.ibt.unam.mx/gecont/	Univ. Nacl. Aut. de México	(Martinez-Guerrero et al. 2008)	Comparative genome analysis
GenomeVx	wolfe.gen.tcd.ie/GenomeVx/	Trinity College, Dublin	(Conant and Wolfe 2008)	Circular chromosome visualization
Gview	www.gview.ca/	Natl. Microbiol. Lab., Winnipeg	(Petkau et al. 2010)	Circular and linear viewer for prokaryotic genomes
IMG	img.jgi.doe.gov/w	DOE-JGI	(Markowitz et al. 2012)	Comparative genomics workbench
Map Viewer	www.ncbi.nlm.nih.gov/mapview/	NCBI	(Wolfsberg 2010)	Viewer to browse genomic sequence
MGV	mgv2.cmbi.ru.nl	CMBI, RUNijmegen MC	(Kerkhoven et al. 2004)	Visualization for prokaryote comparative genomics
MicrobesOnline	www.microbesonline.org	CSHL	(Dehal et al. 2010)	portal for comparative and functional genomics
Microscope	www.genoscope.cns.fr/agc/microscope/	Genoscope	(Vallenet et al. 2009)	Comparative genomics workbench
PSAT	www.nwrce.org/psat	Univ. Washington	(Fong et al. 2008)	a web tool to compare genomic neighborhoods of multiple prokaryotic genomes
SEED viewer	www.theseed.org	Fellowship for Interpretation of Genomes, Burr Ridge	(Overbeek et al. 2005)	Genome viewer connected to the SEED annotation database
UCSC Archaeal Genome Browser	archaea.ucsc.edu/genomes/bacteria/	UCSC	(Chan et al. 2012)	server to browse genomesequence data
WebGBrowse	webgbrowse.cgb.indiana.edu/	UCSC	(Podicheti and Dong 2010)	visualize genome annotation on GBrowse using webserver

TABLE 12. Tools for orthology detection

Tool	Website	Institute	Reference	Content
Berkeley PHOG	Phylofacts.berkeley.edu/orthologs	UC Berkeley	(Datta et al. 2009)	Orthology identification plus Database
COG	www.ncbi.nlm.nih.gov/COG	NCBI	(Tatusov et al. 2003)	annotated clusters of orthologous sequences
EggNOG	Eggnog.embl.de	EMBL	(Powell et al. 2012)	Database of Orthologous groups
EGO	Compbio.dfci.harvard.edu/tgi/ego	Harvard Univ.		Eukaryotic Gene Orthologs
Inparanoid	www.cgi.ki.se/inparanoid	Karolinska Inst., Stockholm	(Remm, Storm, and Sonnhammer 2001)	Ortholog determinationon basis of biderectionality
LOFT	www.cmbi.ru.nl/LOFT	CMBI, RUNijmegen MC	(van der Heijden et al. 2007)	Orthology prediction at scalable resolution by phylogenetic tree analysis
OrthoInspector	Ebgi.igbmc.fr/orthoinspector	CNRS/INSERM/UDS	(Linard et al. 2011)	Orthology analysis
OrthoMCL	Orthomcl.cbil.upenn.edu	Univ. of Pennsylvania	(Fischer et al. 2011)	Ortholog identification plus database
Roundup 2	Roundup.hms.harvard.edu	Harvard Med. School	(Deluca et al. 2012)	High throughput orthology database

TABLE 13. Annotation pipelines

Pipeline	Website	Institute	Reference	Content
BASys	wishart.biology.ualberta.ca/basys	Univ. Alberta	(Van Domselaar et al. 2005)	web server that supports automated, in-depth annotation of bacterial genomic
3D-Genomics	www.sbg.bio. ic.ac.uk/3dgenomics	Imp. College, London	(Fleming et al. 2004)	a database to compare structural and functional annotations of proteins between sequenced genomes
easyDAS	Ebi.ac.uk/pand-srv/easydas	EMBL-EBI	(Gel Moreno et al. 2011)	Customized annotation pipeline
DAS Writeback	"	"	(Salazar et al. 2011)	Community annotation
Genetools	www.genetools.no	NTNU, Trondheim	(Beisvag et al. 2006)	Annotation of single genes or lists of genes based on content public databases
IMG	img.jgi.doe.gov/w	DOE-JGI	(Markowitz et al. 2012)	Comparative genomics workbench
ISGA	isga.cgb.indiana.edu	Indiana Univ.	(Hemmerich et al. 2010)	Annotation pipeline for prokaryotic genomes
KAAS	www.genome.jp/kegg/kaas/	Kyoto Univ.	(Moriya et al. 2007)	Community annotation and whole genome annotation pipeline
MicrobesOnline	www.microbesonline.org	CSHL	(Dehal et al. 2010)	portal for comparative and functional genomics
Microscope	www.genoscope.cns.fr/agc/microscope/	Genoscope	(Vallenet et al. 2009)	Comparative genomics workbench
RAST	RAST.nmpdr.org	Fellowship for the interpretation of genomes	(Aziz et al. 2008)	Community annotation and whole genome annotation pipeline

TABLE 14. Some Databases with regulatory motifs and Search tools

Tool	Website	Institute	Reference	Content
databases				
DBTBS	dbtbs.hgc.jp	Inst. Med. Sci., Univ. Tokyo	(Sierro et al. 2008)	*Bacillus subtilis*
MicrobesOnline	www.microbesonline.org	CSHL	(Dehal et al. 2010)	portal for comparative and functional genomics
PRODORIC	www.prodoric.de	Techn. Univ. Braunschweig	(Grote et al. 2009)	prokaryotic elements
RegTransBase	regtransbase.lbl.gov	LBNL, IITP	(Kazakov et al. 2007)	transcription factor binding sites
RegPrecise	regprecise.lbl.gov	LBNL, IITP	(Novichkov et al. 2010)	regulon predictions
RegulonDB	regulondb.ccg.unam.mx	CGG/UNAM	(Gama-Castro et al. 2011)	*Escherichia coli*
SCPD	http://rulai.cshl.edu/SCPD/	CSHL	(Zhu and Zhang 1999)	*Sacharomyces cerevisiae*
Transfac	www.gene-regulation.com/pub/databases.html	BioBase GmbH	(Matys et al. 2006)	Eukaryotic regulatory elements
Motif finding tools				
ARCS-motif	beijing.case.edu/ARCS Motif/ARCS Motif	Case Western Reserve Univ.	(Zhang, Su, and Yang 2009)	
Bioprospector	ai.stanford.edu/~xsliu/BioProspector	Stanford Univ., Palo Alto	(Liu, Brutlag, and Liu 2001)	
FitBar	archaea.u-psud.fr/fitbar	CNRS	(Oberto 2010)	
GIMSAN	www.cs.cornell.edu/~ppn3/gimsan	Cornell Univ.	(Ng and Keich 2008)	
GLAM2	acb.qfab.org/acb/glam2	ARC	(Bailey et al. 2009)	
MEME	meme.sdsc.edu/meme/cgi-bin/meme.cgi	NCRR/NBCR	(Bailey et al. 2009)	
MicroFootPrinter	bio.cs.washington.edu/software.html	Univ. Washington	(Neph and Tompa 2006)	
W-AlignACE	www1.spms.ntu.edu.sg/~chenxin/W-AlignACE	Nanyang Techn. Univ., Singapore	(Chen et al. 2008)	

TABLE 15. Other Tools and Portals

	Website	Reference
Look up identifiers		
Batch Entrez	www.ncbi.nlm.nih.gov/sites/batchentrez	
SNAD	veb.lumc.nl/SNAD	
Representation of Text		
WebLogo	weblogo.berkeley.edu/logo.cgi	(Crooks et al. 2004)
Wordle	www.wordle.net	
Portals		
	bioportal.weizmann.ac.il	
	www.agbase.msstate.edu	
	www.jgi.doe.gov	
	www.ncbi.nlm.nih.gov	
Tools		
	www.cbs.dtu.dk/services/	
	www.sanger.ac.uk/resources/software	
	sparks.informatics.iupui.edu	(Glanville et al. 2007)
	phylogenomics.berkeley.edu	

Appendix A: annotation of maltose phosphorylase

p1/4 map homologs in *Lactobacillus plantarum*

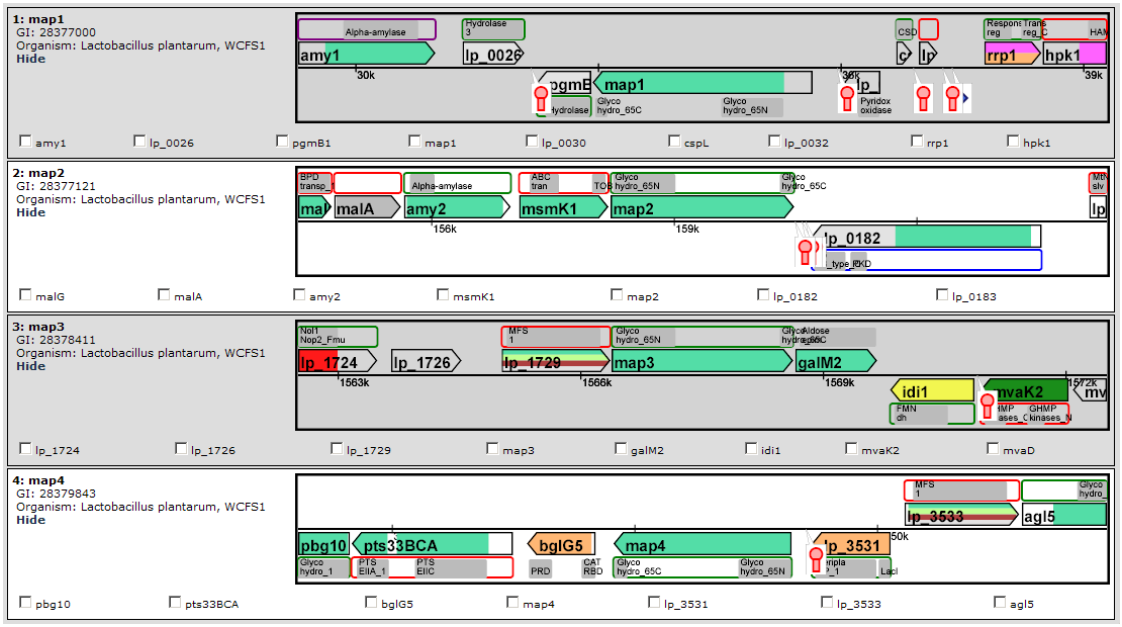
A

FASTA files of sequences extracted from NCBI RefSeq database

```
>28377000_Lactobacillus_plantarum__WCFS1_map1
MKLIRLRIENDAMDIAYHPVSGQAATAHYLIAYNSDQTIGENLENIKVRLAGLQFDAAIL
ENGLSYPFSDTIVGVNYDRIDVGLALTNLLNIPVVSQAAVDQLGLVAAVKAKSAYLKWHL
DYYGQYHGVRNNGQEAMLTIGNGYFGLRGAFLESHADKDNYPGTYVAGVYDQTTTTVHDH
QVKNEDLVNLPNAQFMTFGIDHQTPFTLNEHDLQDAYRSLDLKTGLLTTTKLIQLASGHQ
LRIRSQKVANMRDWHRYSIRYQVTPLNFAGSLQIYTEIDGSVVNSNVSRYNVFDQHHLKT
MGIETAANTVYLSGQTKSSHINYTIGAKLTSPDVPAIENFNSTQQPQGVQQTVSLAVEAG
KTYTFDKNVVIATSNDHSDPQLTHVQAELDQSSFDNTVTTSKDYWEATWRATDIKIRGDI
TSQRLLRVNIYHSFVSAAAIESGQLDASVGARGLHGEAYRGHVFWDEMFILPFYTLHRPE
LAKQLLAYRYRRLPMARKNAEAEGYAGAMYPWQSASKGDEQSQFTHLNPITKTWDPDNSR
LQRHVSLDIAYNVWFYYHVTQDRDFLTHYGMEMLLSIAAFWISKADYDQANGRYNISGVM
GPDEFHENYPNSTTAGLKNNAYTNIMVAWLFDTITTLRAQMPQAAFQAAAQAAQFDSTAE
QALTTISHQLTLEINRRGIIAQFEGYFNLPTLDFQAYRQKYGDIARMDRILKAESKTPDA
YQVAKQADALMAFYNFDVTTVQKIIEKMGYQLPKEYLTHNIQYYLARTTHGSTLSRIVYA
VLNQLDDNDDDAWKLFSEALASDYYDIQGGTTAEGIHLGVMCATLNLTTRNFGGVNPLGA
HLQVNPRLPEHWQTLKFKHLFRGVHYVFVIDHQRVTVTADQPSTIMIGKQHYQLQAKKPL
SVTYTD
>28377121_Lactobacillus_plantarum__WCFS1_map2
MKRIFEVDPWHVISHELVPTDKRLQESMTSIGNGYMGMRGMFEEHYSNDTLKGIYIGGVW
YPDKTRVGWWKNGYPDYFGKVINSVNFIKVNITIDGDQVDLAKDTVSDFTLDLDMHTGIL
RRSFIITKGEKRVQFMIDRFVSVAQKELFDVHYSVHNLGAEPVQIGFISQIDADVFNEDA
NYDEQFWQVLDKSHAITGGSMFAETKPNNFGTPRFTLGMQMLHETAFRGINAPETEKAVT
NIFRGTLAPDETKNFEKRVLVVTSRDYADMAAMRTGLAELSETVSTQSYEDLRQAHVDGW
AKRWQLSDVLIDGDDAAQQGIRFNLFQLFSTYYGEDKRLNLGPKGFTGEKYGGATYWDTE
AFGVPFYLSLAKPEVTENLLEYRYNQLAGAFHNAKMQGLAGALYPMVTFNGIECHNEWEI
TFEEIHRNGSIAYAIFNYTRYTGDEIYLKTKGIDVLTGISRFWADRVHFSERKQQYMIHG
VTGPNEYENNVNNNWYTNFMARWTLEYTLASLKKVDAAKRAALQITDAELAKWQDIVDRM
YFPYDEKLGIFVQQDGFLDKDVKPVSSIPADQRPINQHWSWDHILRSPYIKQADVLQGIY
YFMDKFTAEQKKRNFDFYEPLTVHESSLSPAVHAILAADLHYEDKAVEMYERTARLDLDN
YNNDTVDGLHITSMTGSWLAIVQGFAGMRVKDDTLAFAPFVPKAWTHYQFHVNFRGRLIK
VDVDQQGTTVELVSGEPLTIELNGQAIEVASTVTTK
>28378411_Lactobacillus_plantarum__WCFS1_map3
MKRIFDVNPWHVLTTDFKPENKRLQESMTSLGNGYMGMRGFFEEDYTGDTLPGIYLGGVW
YPDKTRVGWWKNGYPEYFGKVINAVNFIKINFKLNGTKIDLATANFSDFKLDLDMQHGTL
TRSFIVDQDGQRVRVTFERFLSVAQKELSVQKVTFENLSDQAVELKVASALDADVKNEDA
NYDEHFWNVHEVTDDRLIAQTVPNDFGTPQFTSGMQVSYVTSLTHKAAEVTDTATVDRYF
GTLEAGASVSFEKRVVVVTSRDYPTDDAIMTALQRLTDTVSAQSFDDLLAAHEAGWADRW
QQSDIQIEGDTEAQQGMRFNLFELFSTYYGDDPRLNIGPKGFTGEKYGGATYWDTEAFAV
PVYLGITKPEVTRNLLMYRYKQLDGAYHNARQQGLDGALFPMVTFNGIECHNEWEITFEE
IHRNGDIAFAIYNYTRYTGDTSYVLHEGSKVLTEISRFWADRVHFSKRNQQYMIHGVTGP
DEYENNVDNNFNTNYLAQWTLKYTLEILDQVSDDQAAALNVTAEERAHWQDIVDRMYLPY
DKDLDIFVQHDGYLDKDLAPVSAIPADQLPINQHWSWDHILRSPYIKQGDVLQVMYDFID
DFSKTQLKHNFDFYEPMTVHESSLSPAIHAVLAADLHYEDKAVAFYNRTARLDLDNYNND
TVDGLHITSMTGGWIAMVQGFAGMRVHDGQLSYRPFLPKQWTKYSFRQVFRDRIIEVTVD
HDGTTLKLIAGEPIDVQVDGTTQTLTQN
>28379843_Lactobacillus_plantarum__WCFS1_map4
MQIDKLTLKDVATTNRQYLETIFALGDGHMGVRDSVPFTGNQQATLPVMLINGYYATNPI
TYGESAYGYAKNHQTIISLTSPRYLDFATEKTSSEIPGDWDMTVDDAVLDFDTGVLTEKF
QIVTSDDHHFELVVKSMIALDNSHQLFLNYELTSLDYTGTLRFTRPTIHDVTATTADADD
PRVAQRQSGLTNQFIPADRQQQIEWQTLVASTHQQITQRDWLTSFPTGFEVYGDHAGFHG
SGTIEVQTTLSWSFTRQISEINQALPDTITSWTNDIVDRNRQILTNFWQQSQVQISDKKL
QTGIQYNLFQLFQSAGRNGLTNIAAKGITGPGYEGHYFWDTEMYMLPFFIYTQPQIAKQL
LHYRFSILPQARERARSLGVTRGALYAWRTINGEEASAYFPAGTAQYHINADIAHTVKLY
FDVTNDQNFLREQGAAVVLETARFWLQFGSWQQRDGRSQFCLYKVTGPDEYTALVDNNYY
TNRMAKENMAFAAWLLENHYIDGDADEQAQFTKASTNMYLPYDQEQQVTAQDDNSPKMPV
WPFATTAATQYPLLLHYHPLMIYRHRVNKQADTLLAEMLFPEDQSLEQLRRDYEYYEPIT
THDSSLSRSIFSILASRLGDHDKAFSYYMDTSLMDLVDLQGNAKDGLHEANLGGSWLGLT
YGFAGMYVADGKLHITNHLPTQITHLSYRLRFRGRVLEVQLYQDKTQVKLVTGLPLAVVV
AGKEYDVLQGARSE
```

B

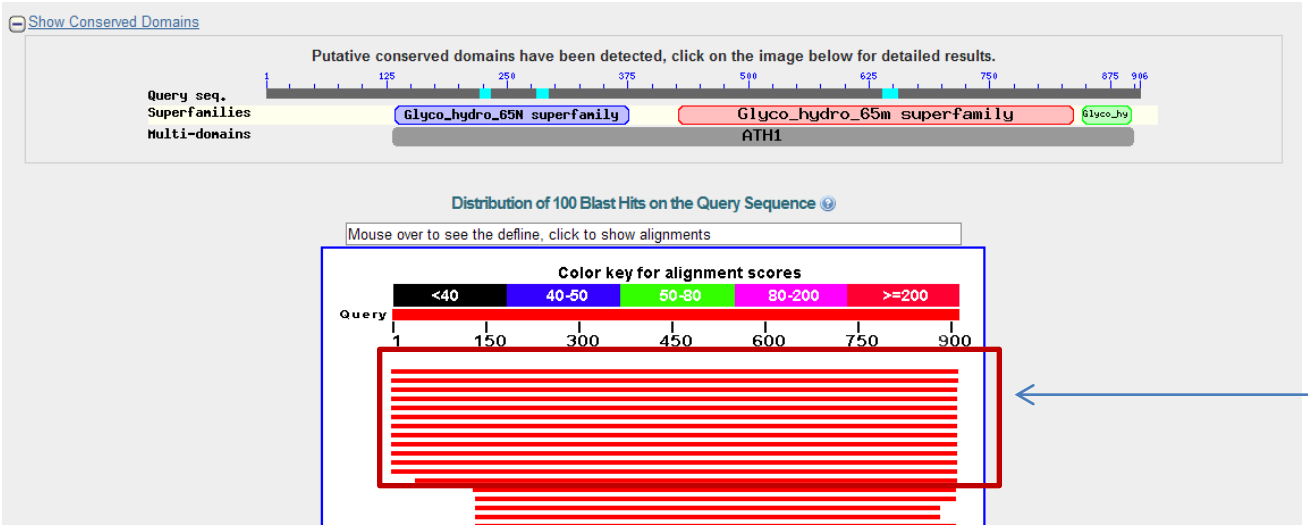
View of the genomic context obtained using the Locus-Tags and the microbial genome viewer http://mgv2.cmbi.ru.nl/



Appendix A: annotation of maltose phosphorylase

p2/4 BLAST results for map1

C



Clear visual separation of orthologs on basis of alignment length and score

Sequences producing significant alignments:

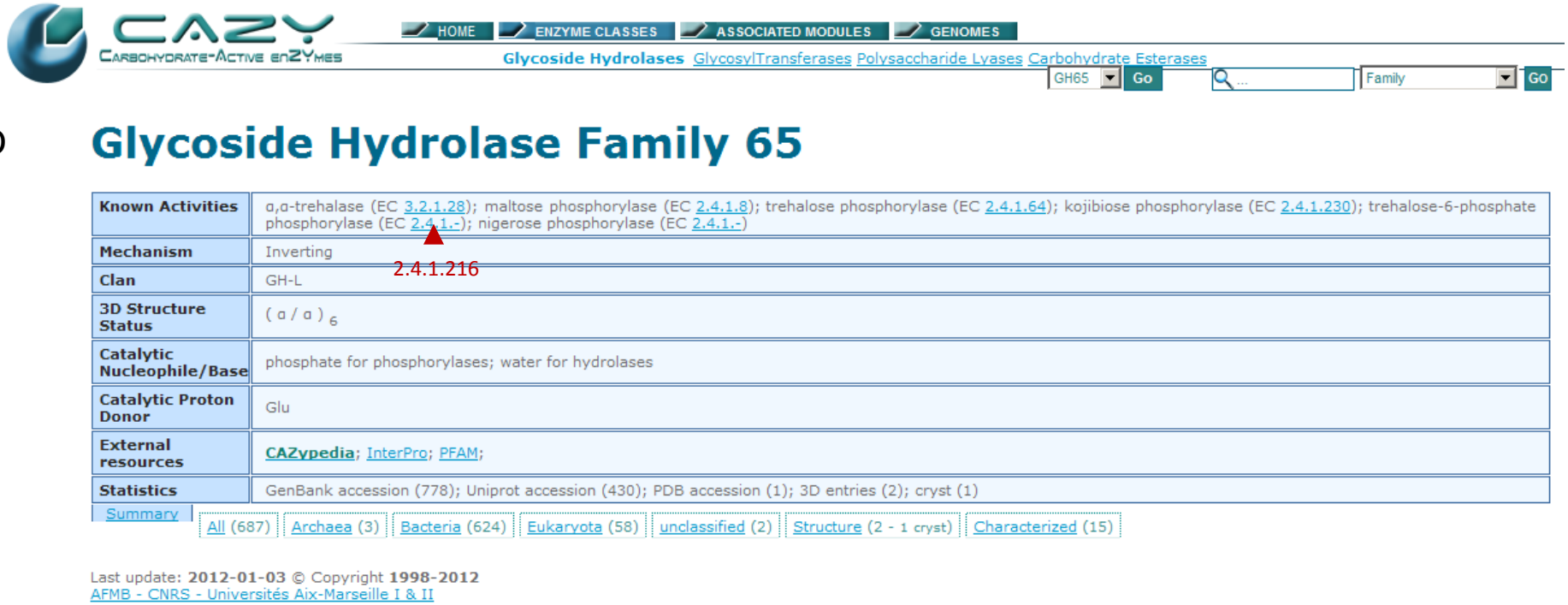
Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NP_783892.1	maltose phosphorylase [Lactobacillus plantarum WCFS1] >gi|254555	1883	1883	100%	0.0	100%	G
ZP_07079020.1	glycosyl hydrolase [Lactobacillus plantarum subsp. plantarum ATCC	1878	1878	100%	0.0	99%	
CCC17049.1	maltose phosphorylase [Lactobacillus pentosus IG1]	1715	1715	99%	0.0	90%	
CCB82026.1	maltose phosphorylase [Lactobacillus pentosus MP-10]	1700	1700	99%	0.0	90%	
ZP_08574699.1	glycosyl hydrolase [Lactobacillus coryniformis subsp. torquens KCTC	1138	1138	99%	0.0	59%	
ZP_08477206.1	glycosyl hydrolase [Lactobacillus coryniformis subsp. coryniformis KC	1130	1130	99%	0.0	59%	
ZP_03954783.1	maltose phosphorylase [Lactobacillus hilgardii ATCC 8290] >gb|EEI2	1085	1085	99%	0.0	55%	
ZP_03943639.1	maltose phosphorylase [Lactobacillus buchneri ATCC 11577] >gb|EE	1083	1083	99%	0.0	55%	
ZP_03940626.1	maltose phosphorylase [Lactobacillus brevis subsp. gravesensis ATC	1082	1082	99%	0.0	55%	
YP_795643.1	trehalose and maltose hydrolase (phosphorylase) [Lactobacillus bre	1078	1078	99%	0.0	57%	G
YP_004398575.1	Trehalose 6-phosphate phosphorylase [Lactobacillus buchneri NRRL I	1076	1076	99%	0.0	56%	G
EHL99058.1	glycosyl hydrolase family 65 central catalytic domain protein [Lactot	1046	1046	99%	0.0	54%	
ZP_08577324.1	Trehalose 6-phosphate phosphorylase [Lactobacillus farciminis KCTC	791	791	95%	0.0	45%	
ZP_08577931.1	glycoside hydrolase family 65 central catalytic [Lactobacillus farcimi	236	236	85%	1e-67	26%	
ZP_08477205.1	Kojibiose phosphorylase [Lactobacillus coryniformis subsp. coryniforn	226	226	84%	1e-64	27%	

Correct function annotation
Erroneous function annotation

Appendix A: annotation of maltose phosphorylase

p3/4 experimentally validated proteins taken from the CAZY reference database http://www.cazy.org/

D



>BAC54904.1|maltose_phosphorylase|Bacillus_sp._RK-1	(Inoue et al. 2002)
>AAV43670.1|maltose_phosphorylase|Lactobacillus_acidophilus	(Nakai et al. 2009)
>Q7SIE1|maltose_phosphorylase|Lactobacillus_brevis	(Huwel et al. 1997; Egloff et al. 2001)
>Q9CID5|Trehalose_6-phosphate_phosphorylase|Lactococcus_lactis	(Andersson, Levander, and Radstrom 2001)
>BAC20640.1|trehalose_phosphorylase|Geobacillus_stearothermophilus	(Inoue et al. 2002)
>Q8L164|Trehalose_phosphorylase|Thermoanaerobacter_brockii	(Maruta et al. 2002)
>Q8L163|Kojibiose_phosphorylase|Thermoanaerobacter_brockii	(Yamamoto et al. 2004)
>ABX42243.1|Nigerose_phosphorylase|Clostridium_phytofermentans_ISDg	(Nihira et al. 2011)

Appendix A: annotation of maltose phosphorylase

E p4/4 BLAST search of all Lactobacillus and Lactococcus genomes

>Q9CID5|Trehalose_6-phosphate_phosphorylase|Lactococcus_lactis (Andersson, Levander, and Radstrom 2001)

>BAC54904.1|maltose_phosphorylase Bacillus_sp._RK-1 (Inoue et al. 2002)
>AAV43670.1|maltose_phosphorylase Lactobacillus_acidophilus (Nakai et al. 2009)
>Q7SIE1|maltose_phosphorylase Lactobacillus_brevis (Huwel et al. 1997; Egloff et al. 2001)

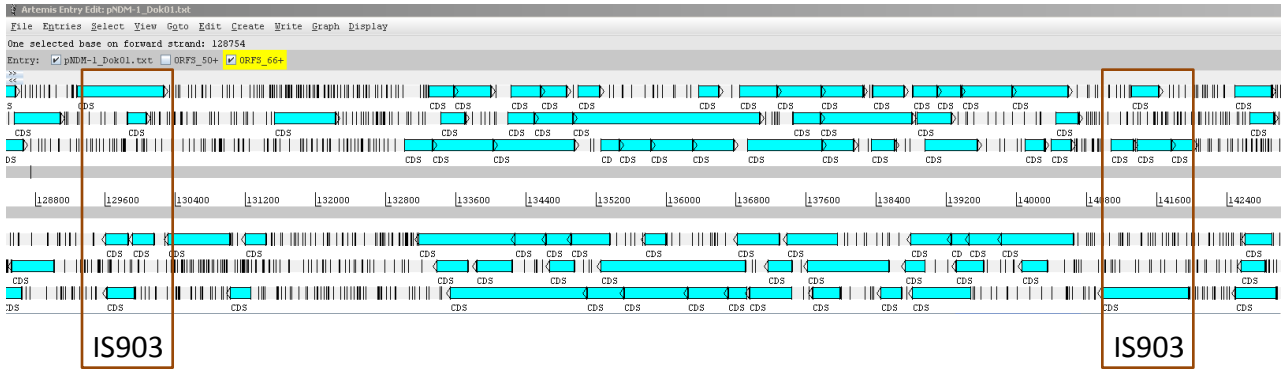
>BAC20640.1|trehalose_phosphorylase|Geobacillus_stearothermophilus (Inoue et al. 2002)
>Q8L164|Trehalose_phosphorylase Thermoanaerobacter_brockii (Maruta et al. 2002)

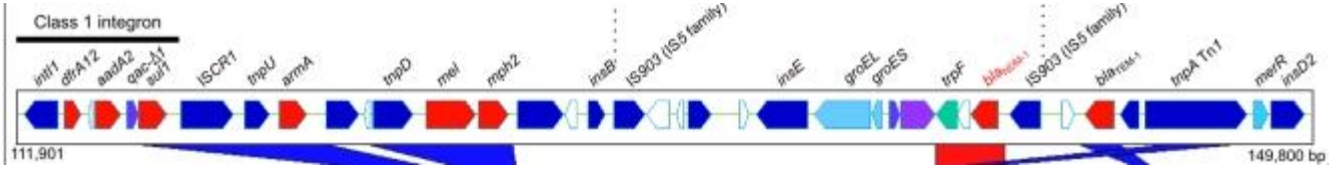
						map1	map2	map3	map4
	00	ref	NP_783892.1	maltose_phosphorylase	Lactobacillus_plantarum_WCFS1	0	1.21E-49	7.75E-49	4.00E-38
	95	ref	YP_003061612.1	maltose_phosphorylase	Lactobacillus_plantarum_JDM1	0	1.21E-49	7.75E-49	4.00E-38
	17	ref	YP_003923345.1	glycosyl_hydrolase	Lactobacillus_plantarum_subsp._plantarum_ST-III	0	2.70E-50	8.22E-49	1.81E-38
	16	ref	YP_795643.1	trehalose_and_maltose_hydrolase_(_phosphorylase)	Lactobacillus_brevis_ATCC_367	0	2.83E-57	2.27E-57	4.62E-32
	16	ref	YP_004398575.1	Trehalose_6-phosphate_phosphorylase	Lactobacillus_buchneri_NRRL_B-30929	0	7.66E-45	1.30E-39	6.24E-30
	90	ref	YP_808506.1	trehalose_and_maltose_hydrolase_(_phosphorylase)	Lactococcus_lactis_subsp._cremoris_SK11	3.81E-166	1.84E-49	4.73E-52	2.29E-38
	22	ref	YP_001031805.1	trehalose/maltose_hydrolase	Lactococcus_lactis_subsp._cremoris_MG1363	5.98E-166	9.12E-50	3.42E-51	3.01E-39
	10	ref	NP_266584.1	hypothetical_protein_L39593	Lactococcus_lactis_subsp._lactis_Il1403	3.28E-165	7.09E-52	4.58E-51	2.94E-38
gi	71	ref	YP_003352951.1	family_65_glycosyl_hydrolase	Lactococcus_lactis_subsp._lactis_KF147	5.73E-164	1.22E-52	1.97E-51	4.61E-38
gi	58338116	ref	YP_194701.1	maltose_phosphorylase	Lactobacillus_acidophilus_NCFM	2.59E-50	0	0	1.15E-67
		f	YP_001578119.1	maltose_phosphorylase	Lactobacillus_helveticus_DPC_4571	2.62E-49	0	0	4.97E-64
		f	YP_004833071.1	maltose_phosphorylase	Lactobacillus_ruminis_ATCC_27782	5.28E-49	0	0	1.78E-64
		f	YP_004563194.1	maltose_phosphorylase	Lactobacillus_kefiranofaciens_ZW3	5.32E-49	0	0	6.49E-65
		f	YP_001987087.1	maltose_phosphorylase	Lactobacillus_casei_BL23	8.63E-49	0	0	2.27E-57
		f	YP_003602367.1	maltose_phosphorylase	Lactobacillus_crispatus_ST1	1.31E-48	0	0	4.53E-65
		f	NP_391337.1	maltose_phosphorylase	Bacillus_subtilis_subsp._subtilis_str._168	3.09E-48	0	0	1.61E-63
		f	YP_004649699.1	family_65_glycosyl_hydrolase	Lactobacillus_reuteri_SD2112	3.12E-48	0	0	1.37E-73
		f	YP_004032750.1	maltose_phosphorylase	Lactobacillus_amylovorus_GRL_1112	3.20E-48	0	0	1.47E-65
		f	YP_004293066.1	maltose_phosphorylase	Lactobacillus_acidophilus_30SC	3.20E-48	0	0	1.47E-65
		f	YP_001842841.1	maltose_phosphorylase	Lactobacillus_fermentum_IFO_3956	4.81E-48	0	0	3.89E-75
		f	YP_806221.1	maltose_phosphorylase	Lactobacillus_casei_ATCC_334	6.61E-48	0	0	2.11E-58
		f	YP_001841048.1	maltose_phosphorylase	Lactobacillus_reuteri_JCM_1112	7.01E-48	0	0	1.88E-73
		f	YP_001270667.1	maltose_phosphorylase	Lactobacillus_reuteri_DSM_20016	7.01E-48	0	0	1.88E-73
		f	YP_536170.1	maltose_phosphorylase	Lactobacillus_salivarius_UCC118	7.62E-48	0	0	1.31E-59
		f	YP_003170691.1	maltose_phosphorylase	Lactobacillus_rhamnosus_GG	8.56E-48	0	0	2.14E-60
		f	NP_784013.1	maltose_phosphorylase	Lactobacillus_plantarum_WCFS1	1.52E-47	0	0	7.05E-66
		f	YP_003923466.1	maltose_phosphorylase	Lactobacillus_plantarum_subsp._plantarum_ST-III	1.52E-47	0	0	7.05E-66
		f	YP_003061756.1	maltose_phosphorylase	Lactobacillus_plantarum_JDM1	1.52E-47	0	0	7.05E-66
		f	YP_003788020.1	maltosephosphorylase	Lactobacillus_casei_str._Zhang	2.29E-47	0	0	2.47E-58
gi	42518299	ref	NP_964229.1	maltose_phosphorylase	Lactobacillus_johnsonii_NCC_533	6.56E-47	0	0	4.93E-66
gi	331702579	ref	YP_004399538.1	Kojibiose_phosphorylase	Lactobacillus_buchneri_NRRL_B-30929	7.43E-47	0	0	1.10E-71
gi	28378411	ref	NP_785303.1	maltose_phosphorylase	Lactobacillus_plantarum_WCFS1	1.14E-46	0	0	6.59E-71
gi	308180566	ref	YP_003924694.1	maltose_phosphorylase	Lactobacillus_plantarum_subsp._plantarum_ST-III	1.14E-46	0	0	6.59E-71
gi	254556620	ref	YP_003063037.1	maltose_phosphorylase	Lactobacillus_plantarum_JDM1	1.14E-46	0	0	6.59E-71
gi	347533887	ref	YP_004840557.1	glycosyl_hydrolase_YvdK	Lactobacillus_sanfranciscensis_TMW_1.1304	3.42E-46	0	0	1.10E-68
gi	281492258	ref	YP_003354238.1	maltose_phosphorylase	Lactococcus_lactis_subsp._lactis_KF147	3.63E-46	0	0	6.50E-62
gi	15673659	ref	NP_267833.1	maltose_phosphorylase	Lactococcus_lactis_subsp._lactis_Il1403	4.54E-46	0	0	6.83E-62
gi	268318784	ref	YP_003292440.1	putative_maltose_phosphorylase	Lactobacillus_johnsonii_FI9785	6.78E-46	0	0	6.65E-66
gi	116628891	ref	YP_814063.1	maltosephosphorylase	Lactobacillus_gasseri_ATCC_33323	9.39E-46	0	0	5.40E-65
gi	258539192	ref	YP_003173691.1	maltose_phosphorylase	Lactobacillus_rhamnosus_Lc_705	1.36E-45	0	0	5.88E-59
gi	125623599	ref	YP_001032082.1	maltose_phosphorylase	Lactococcus_lactis_subsp._cremoris_MG1363	1.82E-45	0	0	5.02E-60
gi	313122942	ref	YP_004033201.1	glycosyl_transferase,_family_65	Lactobacillus_delbrueckii_subsp._bulgaricus_ND02	5.55E-45	0	0	1.13E-63
gi	116512540	ref	YP_811447.1	maltose_phosphorylase	Lactococcus_lactis_subsp._cremoris_SK11	6.09E-45	0	0	7.30E-61
			04778678.1	maltose_phosphorylase	Lactococcus_garvieae_ATCC_49156	2.31E-44	0	0	2.45E-63
			94554.1	maltose_phosphorylase	Lactobacillus_brevis_ATCC_367	3.42E-44	0	0	4.57E-69
			04397794.1	Kojibiose_phosphorylase	Lactobacillus_buchneri_NRRL_B-30929	2.00E-57	6.11E-75	2.06E-73	3.10E-142
			03064401.1	maltose_phosphorylase	Lactobacillus_plantarum_JDM1	4.55E-41	4.81E-70	3.10E-75	0
			03926186.1	maltose_phosphorylase	Lactobacillus_plantarum_subsp._plantarum_ST-III	2.52E-40	2.21E-69	1.21E-75	0
			86735.1	maltose_phosphorylase	Lactobacillus_plantarum_WCFS1	6.52E-40	7.86E-70	4.22E-75	0
			01031837.1	trehalose/maltose_hydrolase	Lactococcus_lactis_subsp._cremoris_MG1363	4.45E-59	2.38E-74	8.69E-84	1.16E-53

Appendix B: BLAST analysis of the pNDM-1_Dok01 plasmid from *Escherichia coli ST38*

Artemis view of the region flanked by the two IS903 elements

```
CDS             127924  128355  c
CDS             128208  128420  c
CDS             128338  128571
CDS             128379  128663
CDS             128421  128636  c
CDS             128531  129007  c
CDS             128561  129085
CDS             129277  130266
CDS             129604  129858  c
CDS             129618  129926  c
CDS             129854  130060
CDS             129907  130155  c
CDS             130312  131016  c
CDS             131025  131255  c
CDS             131197  131433  c
CDS             131531  132220
CDS             133020  133328
CDS             133174  134271  c
CDS             133291  133578
CDS             133329  134027
CDS             133382  133735  c
CDS             133433  133696
CDS             133539  135095  c
CDS             133579  133998
CDS             133838  134233  c
CDS             134028  134954
CDS             134192  134491
CDS             134233  134568
CDS             134272  134625  c
CDS             134492  134929
CDS             134569  134865
CDS             134626  134913  c
CDS             134669  134950  c
CDS             134914  135354  c
CDS             134930  137065
CDS             134995  135243
CDS             135096  135515  c
CDS             135257  136903  c
CDS             135261  135467
CDS             135468  135821
CDS             135516  136247  c
CDS             135763  135996  c
CDS             135822  136304
CDS             136248  136706  c
CDS             136305  136769
CDS             136375  136602
CDS             136707  136916  c
CDS             136816  137289  c
CDS             136840  137265
CDS             136929  137774
CDS             136953  137429  c
CDS             137162  137434  c
CDS             137266  137766
CDS             137380  137952  c
CDS             137456  137761
CDS             137615  138541  c
CDS             137673  137981  c
CDS             137762  138802
CDS             137767  138246
CDS             137775  138137
CDS             138345  138605
CDS             138364  138714
CDS             138453  138683  c
CDS             138728  138949  c
CDS             138796  139254  c
CDS             138813  139472  c
CDS             138817  139092
CDS             138866  139249
CDS             138954  139550
CDS             139093  139359
CDS             139255  139455  c
CDS             139310  139618  c
CDS             139378  139944
CDS             139456  139824  c
CDS             139825  140649  c
CDS             139945  140589
CDS             140063  140350  c
CDS             140100  140312
CDS             140400  140621
CDS             140450  140698
CDS             140970  141959  c
CDS             141081  141329
CDS             141310  141618
CDS             141378  141764
```





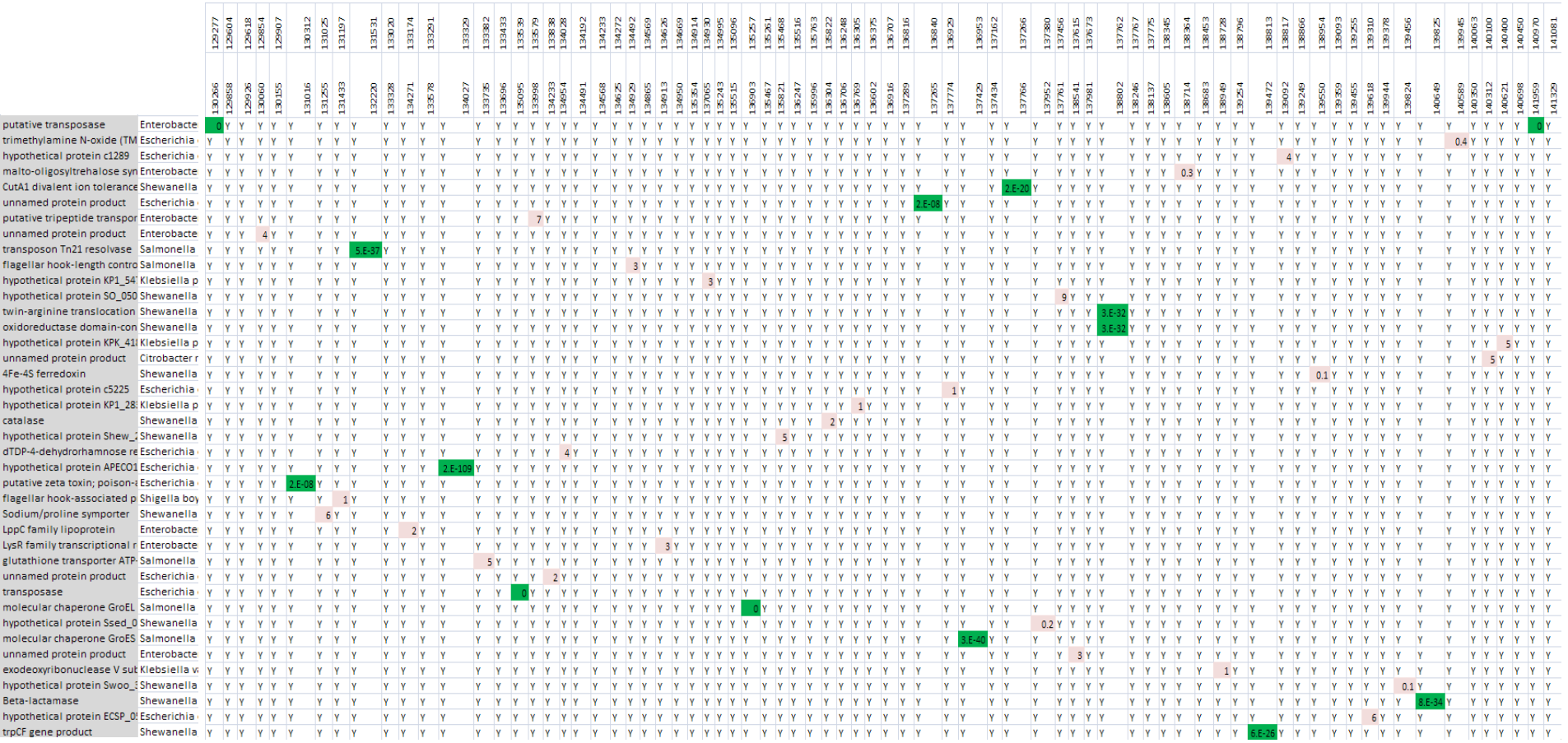
PLoS One. 2011;6(9):e25334. Epub 2011 Sep 23.

Complete sequencing of the bla(NDM-1)-positive IncA/C plasmid from Escherichia coli ST38 isolate suggests a possible origin from plant pathogens.

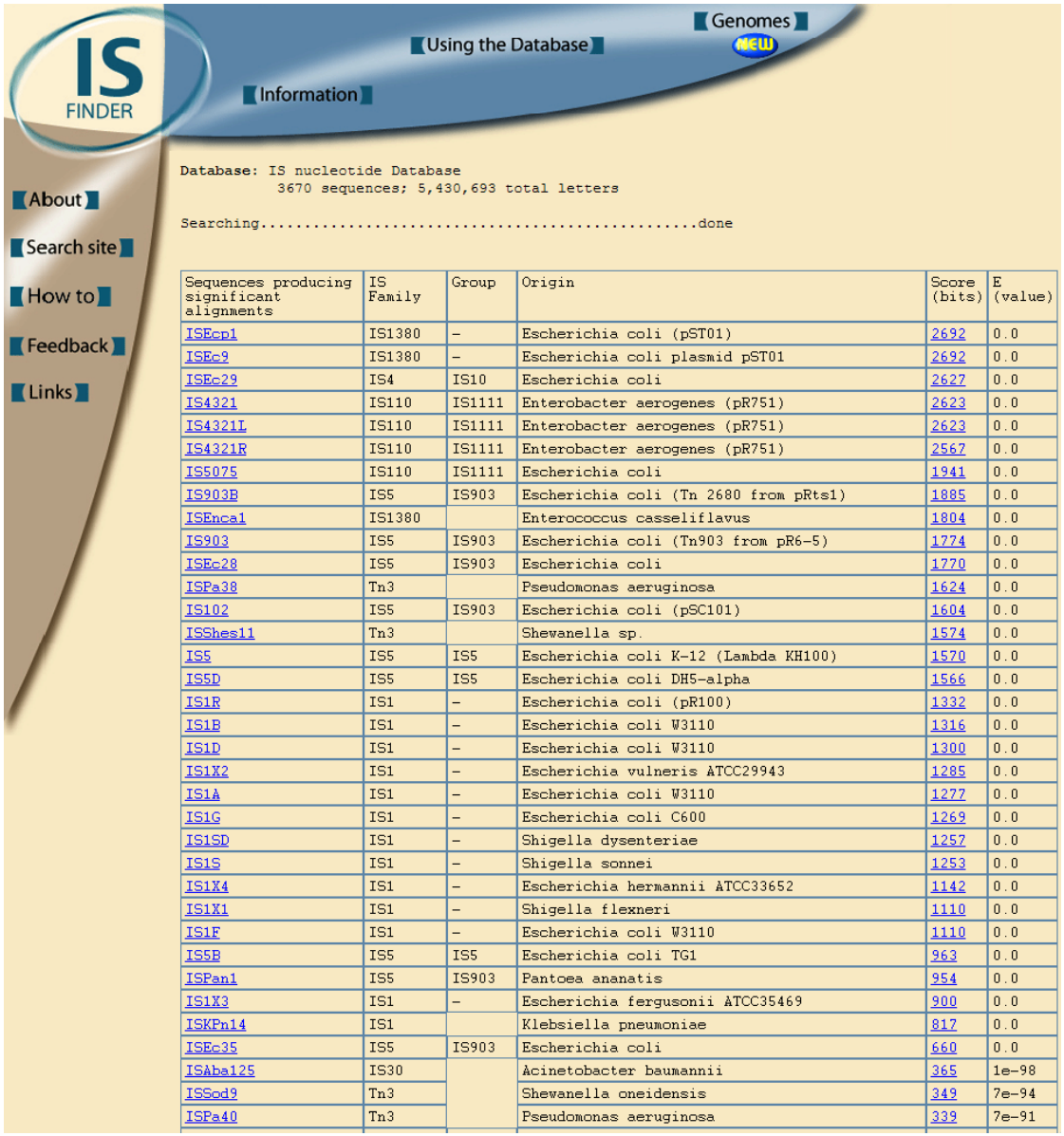
Sekizuka T, Matsui M, Yamane K, Takeuchi F, Ohnishi M, Hishinuma A, Arakawa Y, Kuroda M.

Appendix B: BLAST analysis of the pNDM-1_Dok01 plasmid from *Escherichia coli ST38*

Output of serial BLAST



Appendix B: Identification of protential IS elements on the pNDM-1_Dok01 plasmid from *Escherichia coli ST38*



Database: IS nucleotide Database
3670 sequences; 5,430,693 total letters

Searching.....done

Sequences producing significant alignments	IS Family	Group	Origin	Score (bits)	E (value)
ISEcp1	IS1380	-	Escherichia coli (pST01)	2692	0.0
ISEc9	IS1380	-	Escherichia coli plasmid pST01	2692	0.0
ISEc29	IS4	IS10	Escherichia coli	2627	0.0
IS4321	IS110	IS1111	Enterobacter aerogenes (pR751)	2623	0.0
IS4321L	IS110	IS1111	Enterobacter aerogenes (pR751)	2623	0.0
IS4321R	IS110	IS1111	Enterobacter aerogenes (pR751)	2567	0.0
IS5075	IS110	IS1111	Escherichia coli	1941	0.0
IS903B	IS5	IS903	Escherichia coli (Tn 2680 from pRts1)	1885	0.0
ISEnca1	IS1380		Enterococcus casseliflavus	1804	0.0
IS903	IS5	IS903	Escherichia coli (Tn903 from pR6-5)	1774	0.0
ISEc28	IS5	IS903	Escherichia coli	1770	0.0
ISPa38	Tn3		Pseudomonas aeruginosa	1624	0.0
IS102	IS5	IS903	Escherichia coli (pSC101)	1604	0.0
ISShes11	Tn3		Shewanella sp.	1574	0.0
IS5	IS5	IS5	Escherichia coli K-12 (Lambda KH100)	1570	0.0
IS5D	IS5	IS5	Escherichia coli DH5-alpha	1566	0.0
IS1R	IS1	-	Escherichia coli (pR100)	1332	0.0
IS1B	IS1	-	Escherichia coli W3110	1316	0.0
IS1D	IS1	-	Escherichia coli W3110	1300	0.0
IS1X2	IS1	-	Escherichia vulneris ATCC29943	1285	0.0
IS1A	IS1	-	Escherichia coli W3110	1277	0.0
IS1G	IS1	-	Escherichia coli C600	1269	0.0
IS1SD	IS1	-	Shigella dysenteriae	1257	0.0
IS1S	IS1	-	Shigella sonnei	1253	0.0
IS1X4	IS1	-	Escherichia hermannii ATCC33652	1142	0.0
IS1X1	IS1	-	Shigella flexneri	1110	0.0
IS1F	IS1	-	Escherichia coli W3110	1110	0.0
IS5B	IS5	IS5	Escherichia coli TG1	963	0.0
ISPan1	IS5	IS903	Pantoea ananatis	954	0.0
IS1X3	IS1	-	Escherichia fergusonii ATCC35469	900	0.0
ISKPn14	IS1		Klebsiella pneumoniae	817	0.0
ISEc35	IS5	IS903	Escherichia coli	660	0.0
ISAba125	IS30		Acinetobacter baumannii	365	1e-98
ISSod9	Tn3		Shewanella oneidensis	349	7e-94
ISPa40	Tn3		Pseudomonas aeruginosa	339	7e-91

@ http://www-is.biotoul.fr/