A distributed computational search strategy for the identification of diagnostics targets: Application to finding aptamer targets for methicillin-resistant staphylococci

Keith Flanagan[1], Simon Cockell[2], Colin Harwood[3], Jennifer Hallinan[1], Sirintra Nakjang[3], Beth Lawry[1] and Anil Wipat[1,*]

[1]School of Computing Science, Newcastle University, Newcastle upon Tyne, NE1 7RU

[2]Bioinformatics Support Unit, Newcastle University, Newcastle upon Tyne, NE1 7RU

[3]Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, NE1 7RU

Summary

The rapid and cost-effective identification of bacterial species is crucial, especially for clinical diagnosis and treatment. Peptide aptamers have been shown to be valuable for use as a component of novel, direct detection methods. These small peptides have a number of advantages over antibodies, including greater specificity and longer shelf life. These properties facilitate their use as the detector components of biosensor devices. However, the identification of suitable aptamer targets for particular groups of organisms is challenging. We present a semi-automated processing pipeline for the identification of candidate aptamer targets from whole bacterial genome sequences. The pipeline can be configured to search for protein sequence fragments that uniquely identify a set of strains of interest. The system is also capable of identifying additional organisms that may be of interest due to their possession of protein fragments in common with the initial set. Through the use of Cloud computing technology and distributed databases, our system is capable of scaling with the rapidly growing genome repositories, and consequently of keeping the resulting data sets up-to-date. The system described is also more generically applicable to the discovery of specific targets for other diagnostic approaches such as DNA probes, PCR primers and antibodies.

1 Introduction

The detection of a specific bacterial strain or group of related strains is important in many industrial and clinical settings. For example, in the food industry, early detection of the presence of organisms such as *Salmonella* is needed to prevent contamination [1]. Rapid diagnostic testing is also necessary for the effective detection and treatment of nosocomial or community-acquired infection [2, 3]. Current detection and identification techniques include culture isolation, immunoassays and PCR analyses, but these approaches are expensive and time consuming [4], and most require specialised laboratory facilities. Consequently, these diagnostic tests are of little value for point-of-care screening.

An increasingly attractive alternative to the use of antibodies for the rapid identification of bacterial groups is the use of small molecules, known as aptamers [5]. Aptamers may consist of DNA, RNA, XNA or peptide sequences. Peptide aptamers are particularly promising

[*] To whom correspondence should be addressed. Email: Anil.Wipat@newcastle.ac.uk

doi:10.2390/biecoll-jib-2014-242

because they are relatively easy to construct and handle, and have the ability to bind to proteins with high specificity.

Peptide aptamers consist of a variable peptidic region that is fused at both termini to a scaffold protein [6]. This double fusion limits the conformational liberty of the variable region, increasing the specificity and binding affinity of the aptamer [7]. There is currently considerable interest in developing rapid diagnostic tests which incorporate aptamers to recognise specific molecules. Whilst peptide aptamers typically have specificities similar to antibodies, aptamers have a number of advantages that make them amenable to biosensor applications [8, 9]. The small size of aptamers, typically 20 to 30 amino acids, enables them to reach targets that antibodies may be too large to access [7]. Also, once identified, the highly stable aptamers can be synthesised quickly and cheaply [5]. The high binding specificity of aptamers enables accurate and immediate detection of specific target proteins [10-12].

In order to utilise an aptamer-based sensor, suitable protein targets, or parts of protein targets, must first be identified. A target protein or protein region must be unique to the organism or group of organisms to be detected. For rapid screening it is important that pre-processing of the samples is kept to a minimum. For this reason, the choice of potential target proteins residing in or extending from the cell envelope are the most useful targets, since disruption of the cells in not necessary and the ligands may be used in surface capture-based strategies. Surface-located bacterial proteins are involved in a variety of processes, including host cell targeting, immune evasion, mobility and cell wall metabolism [13]. Surface proteins are also known to facilitate pathogenicity by imparting drug resistance, and are therefore potential candidates for aptamer sensors designed to screen for drug-resistant strains [14, 15].

Recently, the amount of DNA and protein sequence data that is freely available in public databases has grown exponentially. Therefore, a computational approach employing a comparative genomics strategy is now ideally suited for discovering small regions of proteins that are displayed at or near the surface of a cell and are conserved only among members of a defined group of organisms. However, the computational demands of such an approach are immense, given the scale of the data. Recent advances in Grid and Cloud technologies have enabled researchers to access large amounts of computational capacity. Institution-wide Condor [16] installations and commercial services such as Amazon EC2[†] allow access to the processing capacity of many hundreds or thousands of machines, real or virtual. Meanwhile replicated, distributed data stores can utilise the disk and processing resources of several servers simultaneously in order to facilitate the structured storage of billions of data items.

In the work described here, we made use of the large amounts of genomic and proteomic data available in bioinformatics databases, in combination with Cloud computing technology, to discover targets suitable for the detection of bacteria using novel sensor technologies.

Determining suitable diagnostic targets involves a number of steps, which typically form an iterative cycle: a) defining the organism or group of organisms to be detected; referred to here as the group of interest (GOI); b) finding proteins or protein regions that are unique to a GOI; c) determining whether these regions are likely to be surface-accessible; d) assessing whether the structure of the protein is likely to allow an aptamer to bind at the target region(s).

This cycle of steps is typically carried out using a manual approach in which the human user needs to access a variety of information sources, databases and bioinformatics tools. However, given the number of DNA sequences now in the public domain it is no longer possible to carry out the process manually in a systematic and time-efficient manner. Our motivation for this work was the creation of a cloud computing system, ApID, that would

[†] http://aws.amazon.com

perform parts (b) and (c) in an automated fashion, through the creation of workflows, thereby significantly reducing the number of potential aptamer targets while increasing their specificity, and consequently the amount of manual work required in part (d).

We employed a number of bioinformatics applications, including the Cloud computing platform Microbase [17], to develop an automated workflows on which ApID is based. ApID has been designed to identify proteins, or small regions of protein sequences, that are unique to specific groups of organisms. The protein sequence regions identified by ApID are potential aptamer targets since they are common to the defined organisms, but are otherwise globally unique. We demonstrate the utility of the system by identifying aptamer targets unique to the opportunist pathogen, methicillin resistant *Staphylococcus aureus* (MRSA).

2 Methods

The workflows developed for this work made use of the Microbase Cloud computing platform. Microbase is a distributed computing platform that permits the construction of bioinformatics workflows from modular components. Microbase consists of a file storage system, a publish-subscribe messaging system and a distributed job scheduler. Components, termed *responders*, may be activated on receipt of a particular type of event, such as the entry of new data items into the system. Responders can be connected to form a workflow. Each responder is responsible for performing one task, such as executing an analysis application and managing the resulting data. *Actions* within a Microbase pipeline are driven by *events*. For example, the addition of new genome data to a database produces a notification event. The resulting message is then propagated to further responders that are registered to receive this type of message.

A Microbase pipeline may be highly parallel. Responders can be run in parallel, and each responder may generate multiple tasks, each of which, in turn, can run in parallel. For example, a number of independent BLAST tasks may be executed at the same time as a set of protein subcellular localisation prediction tools. The Microbase system, together with responder-specific configuration parameters, determines the exact number of tasks needed, and the computers on which they execute analysis tools or maintain databases of results. Each workflow component was written in Java and executed on a local 48-core cluster, as well as a number of remote Amazon Web Services machines. Microbase managed the distribution and execution of tasks over all the available machines.

The source data used as input to the workflows came from NCBI RefSeq release 47 (2011/05). RefSeq entries for all available bacteria were parsed and stored in a relational database. Result data generated by the workflow was stored in a separate set of PostgreSQL databases. These databases were replicated across a number of high performance machines in order to increase query performance.

3 Results

We developed two automated workflows that form the core of ApID. The first workflow, ApID1, performs data integration and storage, bringing together genomic information including DNA and protein sequences, predictions about protein cellular location and metadata relating to the source of these sequences. A second workflow, ApID2, is then used to query the integrated dataset for surface-located protein sequence fragments that are globally unique to a particular organismal group of interest (GOI).

3.1 Dataset integration and construction workflow (ApID1)

The 'integration and construction' workflow, ApID1 builds the integrated datasets over which ApID2 operates for querying. The dataset construction phase builds upon the output of a number of existing bioinformatics tools. Such tools often operate over independent units of data (jobs), and can therefore be executed in an ‘embarrassingly parallel’ [18] fashion. ApID1 consists of a series of responders that perform the following automated functions: parsing of genome information into a structured database; constructing a protein token database that covers all protein sequences present in the genome database; and the execution of several tools which predict the sub-cellular location of a protein (Figure 1). These protein tokens form the putative target binding regions for the development of diagnostic aptamers.

A set of GenBank-formatted files was used as the primary data source for the data generation pipeline. For this project all of the bacterial genome data from the NCBI RefSeq [19] (accessed 05/2011) was downloaded, parsed and stored in a structured data store, the Genome Pool Database (GPDB). This relational database allows efficient querying over the RefSeq data, and is used extensively by subsequent workflow components.

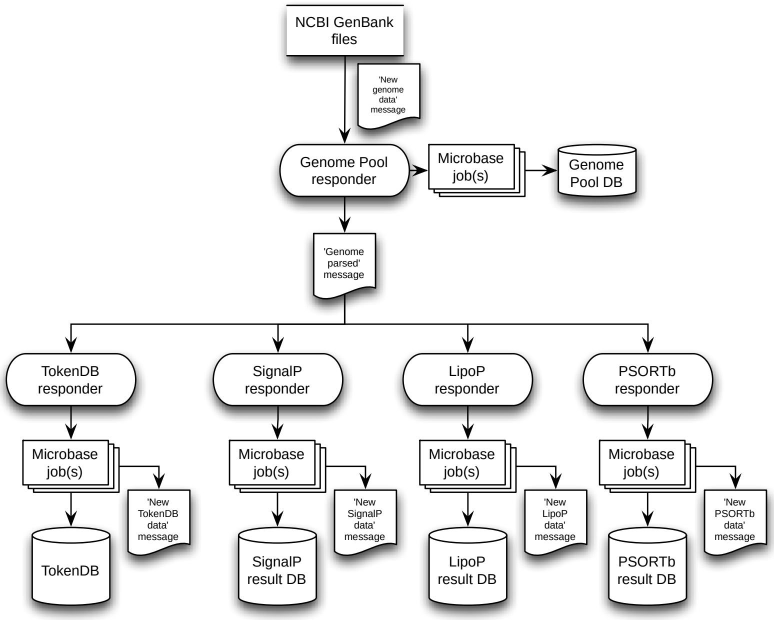

Figure 1: The integration and dataset generation pipeline, ApID1. NCBI GenBank-format files are inserted, parsed and indexed into a database (the GenomePool). This process triggers the execution of a number of analytical tasks that run in parallel: the tokenization of each protein sequence and the sub-cellular localization analysis of each protein. This upstream pipeline is used to generate several datasets that form the basis for further token analysis.

The arrival of a new GenBank fragment into the GPDB triggers the 'protein tokeniser' responder. This pipeline component iterates over all of the protein sequences in the newly added genome fragment. A sliding window is passed over each protein sequence. A token size of 15 amino acids was used with a step size of 1, generating a large number of 15-mers for each protein sequence (Figure 2). The resulting database, TokenDB, stores the set of individual token strings found across all organisms, together with an index to the taxon, protein identifier and intra-protein location for every occurrence of every token string. The tokenisation work was split into Microbase compute jobs; the scope of each job was a single genome fragment, with sizes ranging from a plasmid up to a complete bacterial sequence.

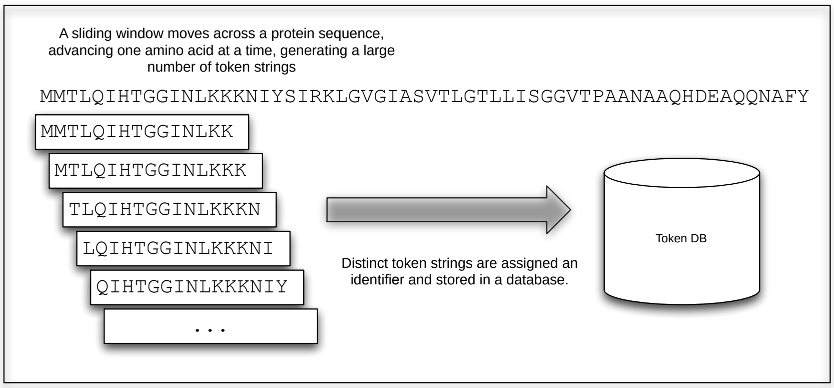
A sliding window moves across a protein sequence, advancing one amino acid at a time, generating a large number of token strings
MMTLQIHTGGINLKKKNIYSIRKLGVGIASVTLGTLLISGGVTPAANAAQHDEAQQNAFY
MMTLQIHTGGINLKK
MTLQIHTGGINLKKK
TLQIHTGGINLKKKN
LQIHTGGINLKKKNI
QIHTGGINLKKKNIY
...
Distinct token strings are assigned an identifier and stored in a database.
Token DB

Figure 2: Generating the protein token database. The individual protein sequences in a GenBank entry are first split into a series of overlapping fragments (tokens). The details of each token, including the position within a particular protein and the taxon from which that protein is derived are then recorded in a database. The tokens are used to specify the binding sites for diagnostic aptamers.

In addition to triggering protein tokenisation tasks, the addition of a new GenBank fragment to the GPDB also triggers the execution of protein sub-cellular localisation tools. TMHMM [20] and SignalP [21] were used to determine which proteins are likely to be surface-associated. Sub-cellular localisation predictions for each protein are stored for future querying. An initial set of 3.7 million bacterial proteins from ~1400 organisms from RefSeq generated 1,349,601,310 tokens. Of these, 888,314,351 tokens were unique.

3.2 Identifying candidate aptamer targets (ApID2)

Once the integrated dataset has been constructed via ApID1, queries can be performed to identify protein sequences, or regions of protein sequences in the form of tokens, that are putative target sequences for a target group of interest (GOI). The process of discovering a putative aptamer targets consists of a number of stages. These stages are defined in the second workflow, the GOI query workflow, ApID2 (Figure 3).

3.2.1 Groups of Interest

Firstly, a GOI must be defined. A GOI consists of a set of taxon identifiers that indicate the organisms that the diagnostic system must be able to distinguish from the background set. The specification of a clearly defined group of organisms is critically important, and defining an optimal GOI is a particularly challenging task. Members of a GOI are typically selected according to user-specified phenotypic features that are unique to a set of organisms. Finding suitable features for diagnostics is problematic for at least two reasons. Firstly, the literature is sometimes ambiguous with regard to the phenotypes of a given bacterial species, and there is no complete reference database of phenotypes for all bacterial species. Secondly, the species concept in bacteria is not well defined, and often phenotypic traits that appear to be attractive targets for diagnostics are not encoded by genes that are maintained by orthologous mechanisms; many of these genes may be acquired through horizontal gene transfer. Therefore, care is needed when selecting candidates for a GOI. If an identified GOI omits organisms with the phenotypic feature of interest the results of the query workflow will be vastly reduced; many potential protein token results will be 'cancelled out' by the sequences of the omitted organism. Conversely, the inclusion in the GOI of even one taxon that does not possess the property of interest reduces the probability that unique tokens will be found.

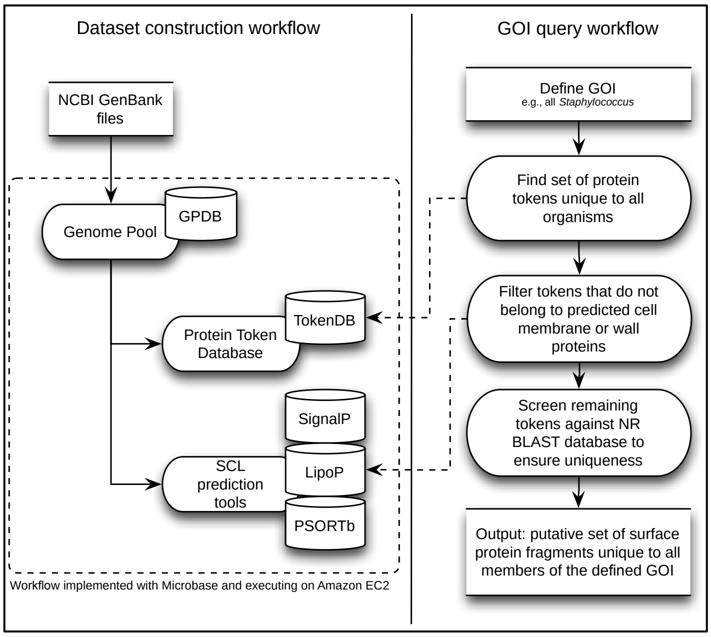
Dataset construction workflow
GOI query workflow
NCBI GenBank files
Genome Pool
GPDB
Protein Token Database
TokenDB
SignalP
SCL prediction tools
LipoP
PSORTb
Workflow implemented with Microbase and executing on Amazon EC2
Define GOI
e.g., all *Staphylococcus*
Find set of protein tokens unique to all organisms
Filter tokens that do not belong to predicted cell membrane or wall proteins
Screen remaining tokens against NR BLAST database to ensure uniqueness
Output: putative set of surface protein fragments unique to all members of the defined GOI

Figure 3: The GOI query workflow, ApID2, uses the datasets provided by the automated token generation and characterisation workflow, ApID1. A Group of Interest (GOI) is defined, based on the phenotype of a set of organisms of interest in a diagnostic context. The TokenDB is queried for any fragments of proteins that occur in all organisms defined in the GOI, but which do not occur in any other organism present in the NCBI RefSeq dataset. The parent proteins of this set of tokens are then identified. Any token that originates in a protein sequence that is not predicted to be located on the membrane or cell wall is excluded. Finally, BLAST-P is used to verify the remaining tokens against the NCBI's non-redundant (NR) database.

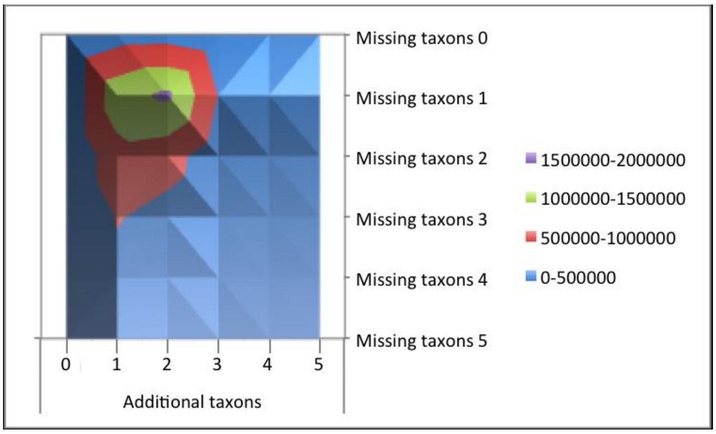

Figure 4: A heatmap showing the distribution of tokens around a Group of Interest. Each cell contains a count of unique diagnostic token strings found by adding or removing organisms from the initial GOI. In this figure, a peak can be seen when 2 extra organisms are added and one of the existing organisms is removed from the original GOI definition. Such a huge increase in tokens may indicate that the original GOI was incorrect, especially if most of the additional tokens found at 2, 1 are from the same two additional organisms.

The establishment of a GOI is therefore an iterative process with a biological domain expert making an initial GOI prediction that may need to be adjusted in the light of available diagnostic signatures. In order to address this uncertainty, the ApID2 workflow provides information allowing a user to explore the 'token space' around a particular group of interest. This functionality allows the user to refine GOI membership through the identification of particular strains that may need to be added or removed from a group. To aid this process a heat map is constructed representing the total numbers of tokens that occur at each possible combination of additional or missing taxa with respect to the GOI (Fig.4). Examination of the results with non-zero values for 'additional' and 'missing' taxa allows the investigator to explore outside of the initial GOI space, by adding taxa to or subtracting taxa from the GOI. For example, if the number of truly unique tokens for a given GOI is relatively small, but a large number of tokens are found when one additional taxon is allowed into the group, then this indicates that the GOI might need to be redefined. Specifically, it indicates that an organism might be missing from the original GOI defined by the user.

3.2.2 Token querying and merging

Once a suitable GOI has been established the purpose of the ApID2 query workflow is to find a set of protein tokens that exist in *all* of the members of the GOI, but that do not exist in *any* other organism present in the database. We refer to these tokens as being *globally unique* to a particular version of the integrated dataset.

A list of token fragments that are unique to the organisms in the GOI is obtained first. The length of this list depends on the number of organisms in the GOI and the extent of their diversity. Typically, a list consists of several million items. All of the token strings in this list occur at least once in one or more members of the GOI. Each token string must then be queried against the entire token database in order to determine in which other proteins it occurs. This step can be performed in a parallel manner if a cluster of database machines is used, but is still a computational bottleneck. If the taxon IDs associated with the proteins returned from each token string query exactly match the set of taxon IDs present in the GOI, then the token string can be considered globally unique to all of the organisms in the GOI.

In many cases, regions larger than a 15-mer token may be common to the members of a GOI, for example when part of a protein is conserved among all organisms in the group. In these cases, it is desirable to merge consecutive, overlapping tokens together into 'super tokens' in order to reduce the analysis workloads and storage space. Variable length unique super tokens, rather than the original 15-mer tokens, are used for the remainder of the sequence selection process.

Whilst each super token is globally unique to the set of organisms in a GOI, it is also desirable to identify sequences similar, but not identical to super tokens in the next closest sequence outside of the GOI. This procedure is necessary since it is possible that an aptamer may be tolerant of mismatches in its target sequence, leading to false positive hits with respect to a given GOI. Furthermore, other ligand-based diagnostics such as antibodies can also be tolerant of amino-acid changes in their epitopes. Therefore, the next stage of the ApID2 workflows involves performing a BLAST similarity search of each super token against the NCBI's non-redundant (NR) database. The super tokens with high similarity to regions of proteins belonging to organisms outside the GOI are discarded.

The final stage of the selection process involves ranking the tokens by the likelihood that they are accessible on the surface of the cell. The predictions from various sub-cellular localisation tools are used to annotate the super-token fragments with tags that indicate whether a particular fragment might be localised on the surface of the cell. These annotations can be visualised within a graphical viewer to browse the available super tokens and either filter or sort them according to the presence or absence of a tag, or the score of a subcellular localisation prediction tool. Intermediate results are stored throughout the selection process, allowing manual inspection and intervention at any point in the pipeline. Once identified, tokens are verified using a series of manual bioinformatics analysis. These analytical steps include verification using a variety of sequence searches against the databases (Blast, FastA etc.), phylogenetic analysis, and the mapping of tokens onto a protein structure, or predicted protein structure, in order to verify that an identified token is actually surface-accessible.

3.3 Use case: Application of ApID to the design of aptamers for the detection of methicillin-resistant *Staphylococcus aureus*.

The ApID system described above was used to find tokens acting as putative targets for Aptamers that would bind uniquely to methicillin-resistant staphylococci. Firstly, a list of all staphylococci with genome sequences in RefSeq (as of 2011/05), together with their methicillin resistance status, was gathered from the literature (Table 1). We then defined a number of GOIs based on the drug resistance information available for the various strains listed in Table 1, and these GOIs were processed by the ApID2 pipeline (Table 2). Three GOIs were identified. Taxon group (a) consists of all *Staphylococcus* proteomes identified as finished in RefSeq. In this GOI, just 22 proteins (one per taxon) had a positive `SpII Lipop` prediction indicating surface localisation. A large number of membrane-associated proteins were predicted using both `psortb` and `TmHMM`. Taxon group (b) consists of just the *Staphylococcus aureus* strains. As expected when comparing a set of more closely related strains, a much larger set of unique token strings was found. On average, in this group, 1900 proteins from each strain shared at least one distinct token string. Many more cell membrane and cell wall proteins were predicted to be suitable aptamer targets. Furthermore, many more `Lipop SpII` predictions are present in GOI (b) than in group (a), and `psortb` predicted 427 cell wall proteins.

Table 1: A summary of the known methicillin resistance/sensitivity of the Staphylococcal strains whose genomes are available in the NCBI RefSeq database. A value of 'ambiguous' in the 'resistance' column means that the literature is contradictory; 'unknown' indicates that no information about methicillin resistance could be found for a particular strain.

Strain	**RefSeq** **Taxon ID**	**Resistant to methicillin**	**Reference**
Staphylococcus aureus RF122	273036	No	[22]
Staphylococcus aureus subsp. *aureus* COL	93062	Yes	[23]
Staphylococcus aureus subsp. *aureus* ED98	681288	No	[22]
Staphylococcus aureus subsp. *aureus* JH1	359787	Yes	[22]
Staphylococcus aureus subsp. *aureus* JH9	359786	Yes	[24]
Staphylococcus aureus subsp. *aureus* MRSA252	282458	Yes	[23].
Staphylococcus aureus subsp. *aureus* MSSA476	282459	No	[23]
Staphylococcus aureus subsp. *aureus* Mu3	418127	Yes	[24]
Staphylococcus aureus subsp. *aureus* Mu50	158878	Yes	[23]
Staphylococcus aureus subsp. *aureus* MW2	196620	Yes	[23].
Staphylococcus aureus subsp. *aureus* N315	158879	Yes	[23]
Staphylococcus aureus subsp. *aureus* NCTC 8325	93061	No	[23]
Staphylococcus aureus subsp. *aureus* str. Newman	426430	No	[22]
Staphylococcus aureus subsp. *aureus* USA300_FPR3757	451515	Yes	[25, 26]
Staphylococcus aureus subsp. *aureus* USA300_TCH1516	451516	Yes	[26, 27]
Staphylococcus carnosus subsp. *carnosus* TM300	396513	No	[28, 29]
Staphylococcus epidermidis ATCC 12228	176280	Ambiguous	[30, 31]
Staphylococcus epidermidis RP62A	176279	Yes	[32]
Staphylococcus haemolyticus JCSC1435	279808	Yes	[31]
Staphylococcus lugdunensis HKU09-01	698737	Unknown	
Staphylococcus pseudintermedius HKU10-03	937773	Unknown	
Staphylococcus saprophyticus subsp. *saprophyticus* ATCC 15305	342451	Unknown	

Group (c) was defined based on all *Staphylococcus* strains for which any evidence of resistance to methicillin was found (Table 1). Only 12 distinct token strings were found, a figure much lower than might be expected. These tokens were located in a single conserved uncharacterised protein, SACOL0037. An investigation of the token space around this GOI was carried out to shed light on the reason for this reduced number of tokens (Figure 5). An analysis of the number of distinct tokens contributed by each strain in a leave one out strategy revealed that *Staphylococcus epidermidis* ATCC 12228 is an outlier to the group. In this strategy, sub-GOIs are created by systematically leaving out one member of the group at a time and counting the number of unique tokens available for that subgroup. Strains whose inclusion results in a major reduction of the unique tokens are highlighted as warranting further investigation into the validity of their inclusion in the GOI. The leave-one-out exercise resulted in the identification of approximately 1,000 distinct tokens contributed by each strain considered, except for *S. aureus* ATCC12228, which provided almost no new tokens. This finding indicates that ATCC12228 may not belong in the original GOI.

With the above results in mind, a further GOI, (d), was defined consisting of any *Staphylococcus* strain for which evidence of methicillin resistance exists (same as group c), but with the exclusion of strain *Staphylococcus epidermidis* ATCC 12228. This strain was excluded from group (d) because of conflicting reports about its resistance to methicillin [30, 31]. This grouping resulted in the prediction of many more distinct tokens (993) from four different sets of homologous proteins. None of these proteins were predicted to possess signal peptides, and none were predicted to be cell wall associated. Twelve proteins, one per organism, were predicted to be membrane associated. This group of 12 was comprised of orthologous members of the well-known penicillin-binding protein 2’, MecA family, SACOL0033. The only other proteins identified in group (d) were MaoC (SACOL0032), IS431mec (transposase) (SACOL0028), and glycerophosphoryl diester phosphodiesterase (SACOL0031). The details of the various GOIs are listed in Table 2.

Table 2: A number of groups of interest (GOIs) were defined (a-d), which were analysed by the analysis pipeline. The number of distinct tokens for each group is shown, together with a summary from various SCL prediction tools for the set of proteins for which at least one distinct token string was present. The GOIs were: a) a group containing all organisms listed in Table 1; b) a subgroup containing just the *S. aureus* strains; c) a subgroup containing organisms for which any evidence of methicillin resistance existed; d) an alternative methicillin resistance group, excluding the ATCC 2228 strain that Takeuchi et al report as methicillin-sensitive.

	GOI			
	(a)	**(b)**	**(c)**	**(d)**
No. organisms	22	15	13	12
No. group-unique token occurrences	97,995	4,625,838	192	11,976
Distinct group-unique token strings	4,438	307,102	12	993
No. proteins containing at least one GOI-unique token instance	10,115	28,593	17 (1 set of homologues)	52 (4 sets of homologues)
SignalP positive predictions	113	1,926	0	0
LipoP SpII positive predictions	22	829	0	0
psortb predictions (Membrane / Wall)	1935 / 22	7791 / 427	0 / 0 (all ‘unknown’)	12 / 0
TmHMM (proteins with > 0 transmembrane domains)	1730	7761	0	12

3.4 Conclusions and future work

As the utility, range and popularity of molecular diagnostics increases, databases of bacterial genome sequences will become an increasingly valuable resource. In this work, we demonstrate how the mining of bacterial genome sequences can be used to design target protein sequences for diagnostic protein aptamers, using the ApID system. Whilst the work we present is specific for protein aptamers, the approach is more generically applicable to any ligand-based diagnostic system where a unique protein sequence target may be used to design an epitope. These systems include antibodies, antibody fragments and, by replacing protein sequences with their encoding DNA sequence, DNA and RNA probes.

We show how the system can be used to specify targets for aptamers specific for methicillin *Staphylococcus aureus*. Aptamers targeting these sequences are being characterised in our laboratory. The ApID system is generically applicable for finding diagnostic protein signatures for a range of applications of relevance to medical care, the food industry and to the environment.

A major requirement of ApID is that the user is able to define a GOI that includes the organisms they would like to detect. The definition of a GOI is a challenging exercise, requiring a combination of user expertise, information about the phenotypic traits of organisms from the literature and databases, and a clear definition of bacterial species. Sources of information relating to these requirements are often prone to error, and the species defined by classical numeric taxonomy frequently do not correlate exactly with those defined through similarity at the genome sequence level. The ApID system therefore allows a user to suggest a GOI and then explore the effect, on the number of unique tokens generated, of modifying the group membership, thus optimising the GOI. In the future it may be possible automatically predefine suggested groups of interest based on the occurrence of shared tokens, saving time by avoiding the need for the dynamic execution of the ApID2 workflow.

A major challenge of the approach implemented in the ApID system is the computationally intensive nature of both the preparation of the integrated datasets and data mining for tokens for a given GOI. Here, we show how the combined use of Cloud computing workflows and parallel databases can be used to address this challenge. We used a system previously constructed in our group, Microbase, to implement the two major workflows underlying the ApID system. Microbase automatically handles the issues of task scheduling, parallelisation and job monitoring and execution. The use of the tokenisation strategy allows a fine-grained definition of diagnostic targets but also adds a layer computational demand. We address this demand through the use of a high performance cluster of replicated databases.

The use of a Cloud-based approach allows the system to be scaled with the number of complete bacterial genome sequences, which continues to increase in an exponential fashion. Furthermore, the use of workflows and the Microbase system allows much of the system to be automated, automatically executing workflows as the databases are updated. In the future we envisage that the system will be enhanced to further exploit the increasingly large number of incomplete genomic sequences that are accumulating in the sequence databases.

Acknowledgements

We would like to acknowledge the assistance of the Newcastle University Bioinformatics Support Unit. This work was supported by EPSRC project EP/G061394/1.

References

[1] M. Taban, S. Aytac, N. Akkoc, and M. Akcelik. Characterization of antibiotic resistance in *Salmonella enterica* isolates determined from ready-to-eat (RTE) salad vegetables. *Brazilian Journal of Microbiology,* 44:385 - 391, 2013.

[2] F. Barbut, L. Surgers, C. Eckert, B. Visseaux, M. Cuingnet, C. Mesquita, N. Pradier, A. Thiriez, N. Ait-Ammar, and A. Aifaoui. Does a rapid diagnosis of *Clostridium difficile* infection impact on quality of patient management? *Clinical Microbiology and Infection,* 20:136-144, 2014.

[3] O. Clerc, G. Prod'hom, L. Senn, K. Jaton, G. Zanetti, T. Calandra, and G. Greub. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry and PCR-based rapid diagnosis of *Staphylococcus aureus* bacteraemia. *Clinical Microbiology and Infection,* 2013.

[4] H. Noorani, E. Adams, S. Glick, S. Weber, S. Belinson, and N. Aronson. Screening for Methicillin-Resistant *Staphylococcus aureus* (MRSA): Future Research Needs: Identification of Future Research Needs From Comparative Effectiveness Review No. 102 *Rockville (MD): Agency for Healthcare Research and Quality (US) Future Research Needs Papers,* 40:2013.

[5] K.-M. Song, S. Lee, and C. Ban. Aptamers and their biological applications. *Sensors,* 12:612-631, 2012.

[6] J. Thibaut, Y. Mérieux, D. Rigal, and G. Gillet. A novel assay for the detection of anti-human platelet antigen antibodies (HPA-1a) based on peptide aptamer technology. *Haematologica,* 97:696-704, 2012.

[7] I. C. Baines and P. Colas. Peptide aptamers as guides for small-molecule drug discovery. *Drug Discovery Today,* 11:334-341, 2006.

[8] L. K. J. Stadler, T. Hoffmann, D. C. Tomlinson, Q. Song, T. Lee, M. Busby, Y. Nyathi, E. Gendra, C. Tiede, and K. Flanagan. Structure−function studies of an engineered scaffold protein derived from Stefin A. II: Development and applications of the SQT variant. *Protein Engineering Design and Selection,* 24:751-763, 2011.

[9] J.-O. Lee, H.-M. So, E.-K. Jeon, H. Chang, K. Won, and Y. Kim. Aptamers as molecular recognition elements for electrical nanobiosensors. *Analytical and Bioanalytical Chemistry,* 390:1023-1032, 2008.

[10] D. Evans, S. Johnson, S. Laurenson, G. Davies, P. Ko Ferrigno, and C. Walti. Electrical protein detection in cell lysates using high-density peptide-aptamer microarrays.

[11] P. Estrela, D. Paul, Q. Song, L. K. J. Stadler, L. Wang, E. Huq, J. J. Davis, P. K. Ferrigno, and P. Migliorato. Label-Free Sub-picomolar Protein Detection with Field-Effect Transistors. *Analytical Chemistry,* 82:3531-3536, 2010.

[12] L.-Q. Gu and J. W. Shim. Single molecule sensing by nanopores and nanopore devices. *Analyst,* 135:441-451, 2010.

[13] W. W. Navarre and O. Schneewind. Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiology and Molecular Biology Reviews,* 63:174-229, 1999.

[14] G. K. Paterson and T. J. Mitchell. The biology of Gram-positive sortase enzymes. *Trends in Microbiology,* 12:89-95, 2004.

[15] A. M. Edwards, J. R. Potts, E. Josefsson, and R. C. Massey. *Staphylococcus aureus* host cell invasion and virulence in sepsis is facilitated by the multiple repeats within FnBPA. *PLoS Pathogens,* 6:e1000964, 2010.

[16] M. J. Litzkow, M. Livny, and M. W. Mutka. Condor-a hunter of idle workstations. in *Distributed Computing Systems, 1988., 8th International Conference on*, 1988, pp. 104-111.

[17] K. Flanagan, S. Nakjang, J. Hallinan, C. Harwood, R. P. Hirt, M. R. Pocock, and A. Wipat. Microbase2.0: A generic framework for computationally intensive bioinformatics workflows in the cloud. presented at the 2012 International Symposium on Integrative Bioinformatics (IB2012), Hangzhou, China, 2012.

[18] C. Moler. *Matrix Computation on Distributed Memory Multiprocessors*. Philadelphia: Society for Industrial and Applied Mathematics, 1986.

[19] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research,* 40:D130-D135, 2012.

[20] A. Krogh, B. Larsson, G. Von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology,* 305:567-580, 2001.

[21] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods,* 8:785-786, 2011.

[22] H. Xue, H. Lu, and X. Zhao. Sequence diversities of serine-aspartate repeat genes among *Staphylococcus aureus* isolates from different hosts presumably by horizontal gene transfer. *PLoS ONE,* 6:e20332:2011.

[23] K. Hardy, D. Ussery, B. Oppenheim, and P. Hawkey. Distribution and characterization of staphylococcal interspersed repeat units (SIRUs) and potential use for strain differentiation. *Microbiology,* 150:4045-4052, 2004.

[24] I. A. Al-Zahrani, C. Hamson, D. Edge, J. Collins, J. D. Perry, M. Raza, G. K., and C. R. Harwood. A *SmaI* restriction site-based multiplex PCR for typing of hospital and community acquired *Staphylococcus aureus*. *Journal of Clinical Microbiology,* 49:3820-3828, 2011.

[25] B. A. Diep, S. R. Gill, R. F. Chang, T. H. Phan, J. H. Chen, M. G. Davidson, F. Lin, J. Lin, H. A. Carleton, and E. F. Mongodin. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. *The Lancet,* 367:731-739, 2006.

[26] S. K. Highlander, K. G. Hultén, X. Qin, H. Jiang, S. Yerrapragada, E. O. Mason, Y. Shang, T. M. Williams, R. M. Fortunov, and Y. Liu. Subtle genetic changes enhance virulence of methicillin resistant and sensitive *Staphylococcus aureus*. *BMC Microbiology,* 7:99, 2007.

[27] K. Nicolaou, N. L. Simmons, J. S. Chen, N. M. Haste, and V. Nizet. Total synthesis and biological evaluation of marinopyrrole A and analogs. *Tetrahedron Letters,* 52:2041-2043, 2011.

[28] W. Tesch, A. Strässle, B. Berger-Bächi, D. O'Hara, P. Reynolds, and F. Kayser. Cloning and expression of methicillin resistance from *Staphylococcus epidermidis* in *Staphylococcus carnosus*. *Antimicrobial Agents and Chemotherapy,* 32:1494-1499, 1988.

[29] R. Rosenstein, C. Nerz, L. Biswas, A. Resch, G. Raddatz, S. C. Schuster, and F. Götz. Genome analysis of the meat starter culture bacterium *Staphylococcus carnosus* TM300. *Applied and Environmental Microbiology*, 75:811-822, 2009.

[30] Y. Q. Zhang, S. X. Ren, H. L. Li, Y. X. Wang, G. Fu, J. Yang, Z. Q. Qin, Y. G. Miao, W. Y. Wang, and R. S. Chen. Genome-based analysis of virulence genes in a non-biofilm-forming *Staphylococcus epidermidis* strain (ATCC 12228). *Molecular Microbiology,* 49:1577-1593, 2003.

[31] F. Takeuchi, S. Watanabe, T. Baba, H. Yuzawa, T. Ito, Y. Morimoto, M. Kuroda, L. Cui, M. Takahashi, and A. Ankai. Whole-genome sequencing of *Staphylococcus haemolyticus* uncovers the extreme plasticity of its genome and the evolution of human-colonizing staphylococcal species. *Journal of Bacteriology,* 187:7292-7308, 2005.

[32] S. R. Gill, D. E. Fouts, G. L. Archer, E. F. Mongodin, R. T. DeBoy, J. Ravel, I. T. Paulsen, J. F. Kolonay, L. Brinkac, and M. Beanan. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. *Journal of Bacteriology*, 187:2426-2438, 2005.