
EuroGentest Unit 2

Information sources and bioinformatics tools

Quality assessment of bioinformatics tools for
genetic testing services

July 2006



EuroGentest



Prepared by:

Andrew Devereau (Andrew.devereau@cmmc.nhs.uk)

Nic Walker (walker@bioinf.man.ac.uk)

National Genetics Reference Laboratory (Manchester)
Central Manchester and Manchester Children's University Hospitals NHS Trust
St Mary's Hospital
Hathersage Road
Manchester
M13 0JH
UK

As part of the activities of:

EuroGentest Unit 2
Unit 2 leader: Segolene Ayme

within the work package WP2.2.

EuroGentest Project Co-ordinator: Professsor Jean-Jaques Cassiman
K.U.Leuven
Center for Human Genetics/ Centrum Menselijke Erfelijkheid
Gasthuisberg O&N
Herestraat 49
Box 602
3000 Leuven
Belgium

The views expressed in this document are those of the work package participants and interviewees and do not necessarily reflect the policies of the institutions or companies they are affiliated to.

© July 2006 Copyright EuroGentest Network of Excellence Project 2005 - EU Contract no. FP6-512148

Table of Contents

GLOSSARY	ii
1 SUMMARY.....	1
2 INTRODUCTION	2
2.1 Project aims and background	2
2.2 Scope	2
3 LITERATURE SURVEY	2
3.1 Development of bioinformatics tools	2
3.2 Categories of tools	3
3.2.1 Databases and data resources	3
3.2.2 Search and analysis tools	6
3.2.3 Interface and visualisation tools.....	11
3.3 Approaches to rating	11
3.3.1 Feature lists	12
3.3.2 Subjective assessment	13
3.3.3 External/Objective assessment	13
3.3.4 Medical guidelines approach.....	14
3.4 Conclusions	14
4 SURVEY OF TESTING LABORATORIES.....	15
4.1 Method	15
4.2 Results.....	16
4.2.1 Survey overview.....	16
4.2.2 Classification of tools discovered	16
4.2.3 Summary of tools discovered.....	17
4.3 Conclusions	18
5 ARCHIVE DEVELOPMENT	18
5.1 Requirements	18
5.2 Design	18
5.3 Implementation	19
5.4 Analysis of Sequence Analysis Tools	21
6 CONCLUSION.....	23
7 BIBLIOGRAPHY	24
8 APPENDIX A.....	27
8.1 Survey responses.....	27

Glossary

Audit trail	A record of the actions on a computer system, e.g. who has used the system, who has changed data values and what the old and new values are.
Benchmark testing	Using a series of standard tests to compare the relative performance of computer systems or software
Bioinformatic tool	A software application used in bioinformatics
Bioinformatics	The use of computational methods in genetics and genomics.
BLAST	Basic Local Alignment Search Tool. A bioinformatics tool that finds regions of local similarity between sequences.
Core database	A genetic mutation database holding data for many genetic loci.
Database	A computer application that allows data to be stored and retrieved. See <i>flat file database</i> and <i>relational database</i> .
EBI	European Bioinformatics Institute. See www.ebi.ac.uk
EMBOSS	European Molecular Biology Open Software Suite. See http://emboss.sourceforge.net/
False negative	For a detection method, a negative result that is incorrect, i.e. it should have been positive. See <i>True Negative/Positive</i> .
False positive	For a detection method, a positive result that is incorrect, i.e. it should have been negative. See <i>True Negative/Positive</i> .
Flat file database	A simple type of database in which data are held as lists stored in text files.
GUI	Graphical User Interface. A userinterface (<i>cf</i>) based on a graphical as opposed to a text-based display.
HGMD	Human Gene Mutation Database. A core database (<i>cf</i>) run by the Institute of Medical Genetics, Cardiff, UK. See www.hgmd.cf.ac.uk
HGVS	Human Genome Variation Society, formerly the HUGO (<i>cf</i>) Mutation Database Initiative. Fosters the discovery and characterization of genomic variations. Has among other initiatives produced guidelines for mutation nomenclature. See www.hgvs.org
HUGO	Human Genome Organisation. The international organisation of scientists involved in human genetics. See www.hugo-international.org/
LSDB	Locus specific database. A genetic mutation database holding data for one or a few related genetic loci, usually with more detail than a core database (<i>cf</i>).
Metadata	'Data about data'. Metadata is used to describe a resource, e.g. the subject, author, title etc. of a web page.
NGRL (Manchester/Wessex)	National Genetics Reference Laboratory at Manchester and Wessex in the UK: two organisations working on behalf of genetic testing centres in England.
OMIM	Online Mendelian Inheritance in Man. An online database of human genes and genetic disorders. See http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Parameter setting	The action of setting optional values on a bioinformatics tool, e.g. gap penalties in a BLAST tool, which affects the way that the tool operates.
Pipeline	A bioinformatics application that passes data between a series of existing bioinformatics tools in turn to achieve an overall specific goal.

Positive predictive value	For a detection method, a measure of the probability that a positive test result is correct. Expressed as the number of true positive results as a percentage of all positive results.
RefSeq	A Reference Sequence collection which aims to provide a comprehensive, integrated, non-redundant set of genetic sequences for major research organisms. See www.ncbi.nlm.nih.gov/RefSeq/
Relational database	A type of database in which data are held in a structured way based on a mathematical model, often visualised as a series of related tables. This approach is more complex and flexible than a flat file database (<i>cf</i>) and is widely used in many fields.
Sensitivity	A measure of the ability of a detection method to identify positive cases correctly. Expressed as the number of true positive (<i>cf</i>) results as a percentage of the number of positive cases tested.
Specificity	A measure of the ability of a detection method to identify negative cases correctly. Given as the number of true negative (<i>cf</i>) results as a percentage of the number of negative cases tested.
Toolkit	A computer tool that allows a component of an application to be developed, e.g. a graphics toolkit allows a GUI (<i>cf</i>) to be developed.
True negative	For a detection method, a negative result that is correct. <i>See False Negative/Positive.</i>
True positive	For a detection method, a positive result that is correct. <i>See False Negative/Positive.</i>
UI	User Interface. The part of a software application that provides the display to the user and allows interaction. Normally a GUI (<i>cf</i>).
Widget	A component of a computer interface such as a window or text box.
Wiki	A software application that provides an interactive web site that allows users to create and edit the contents of pages via a web browser. Used for creation of collaborative resources.
WP	Work Package

1 Summary

Work package 2.2 of the EuroGentest project aims to map the bioinformatics requirements of the network and to determine the quality of these tools and how they might be validated for their use in diagnostics. This report represents the results of a literature survey and laboratory survey designed to determine tools used in genetic testing, and work to develop proposals for the assessment of quality and a prototype tool archive.

The literature survey of bioinformatics tools highlighted two aspects of quality that should be developed. The first is provision of information about the tool: what it does, who it is aimed at, who developed it, how it will be supported, its scope etc., and how to use it. Provision of this information is often mixed and unstructured, but is important for the correct choice and use of tools and to record the provenance of data. It is suggested that three areas are addressed: information about tools, information about data, and information about the operation of the tool.

The second aspect is the assessment of the performance of a tool, such as its sensitivity or specificity, or the completeness of its data coverage. This information is vital for the accurate interpretation and reporting of results.

It is proposed that the first aspect can be addressed by categorising tools according to their purpose and developing a list of features of importance for each tool type. Information can then be added to individual tools to provide guidance on their operation. The second aspect requires standardised tests to be developed for performance parameters specific to each tool type. It is suggested that the development and operation of such systems will require an independent body.

A survey of testing centres was undertaken to discover which tools are being used. Eleven centres in the EU were interviewed, including representatives of molecular, biochemical and cytogenetic laboratories. Most of the software found was commercial software associated with particular pieces of equipment or processes, and databases. The proposal to categorise and describe tools and devise methods for their performance testing therefore proceeded with the tool types found, with Sequence Analysis tools the first category to be addressed.

A prototype web-based archive of tools has been established using a 'wiki,' i.e. a web site that allows collaborative documents to be developed by a group of users via web browsers. The tool types found from the laboratory survey have been classified and the Sequence Analysis tools further analysed to produce a table of specific features that are important for this tool type. A Technology Assessment (Patel and Wallace, 2005b) conducted at NGRL(Manchester) for one sequence analysis package has been used to enter data into the table and to propose quality assessment measures. The latter are based on the tool's reported and measured sensitivity and its positive predictive value. It is proposed that other tools in this category are assessed using the same dataset in order to provide comparable results.

It is concluded that expert groups for each tool type must be formed to propose and review the quality features and performance measures proposed. This is a priority now for Sequence Analysis tools, but database tools are proposed as the next area of attention.

The archive is being established at ngri.man.ac.uk/mediawiki.

2 Introduction

2.1 Project aims and background

This paper presents proposed approaches to the quality assessment of bioinformatics tools used by the genetics testing network in Europe. It forms part of work package 2.2 within Unit 2 of the EuroGentest project, which seeks to map the bioinformatics requirements of the genetic testing network and then to determine the quality of these tools and propose how they might be validated for use in diagnostics.

Relevant data from a literature survey of tools and their categorisation are summarised and proposals for applying these methods to the categories of tools found are made. This is followed by the results of a survey made within European genetic testing laboratories. Finally the design and implementation of an initial system for archiving and describing the quality of tools is presented.

2.2 Scope

EuroGentest encompasses those disciplines involved in genetic testing, i.e. molecular genetics, cytogenetics and biochemical genetics. In the project we intend to survey the use of bioinformatics tools in representative laboratories throughout Europe. Which computer-based tools constitute regularly used bioinformatics applications was one of the questions raised during the survey.

3 Literature survey

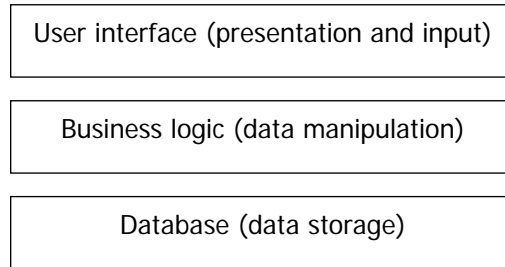
3.1 Development of bioinformatics tools

Reviews of the development of bioinformatics (Hagen 2000; Searls 2000; Ouzounis and Valencia 2003) suggest that techniques developed at the advent of computational biology in the 1970s still influence bioinformatics: analysis tools for sequence alignment, comparison of similar and homologous proteins and protein structure analysis were among the first tools to be developed; followed by development of databases for storage of sequences and results; and visualisation tools for presentation of data.

Searls (2000) suggests the following classification for bioinformatics tools:

- Databases and Data Resources
 - Database technology
 - Public databases
 - Web resources
- Search and Analysis Tools
 - Similarity search and alignment
 - Pattern discovery and search
 - Gene finding
 - Gene expression
 - Genome annotation
 - Other tools
- Interfaces and visualisation tools.
 - Graphical User Interfaces
 - Scientific visualisation

I have adopted these categories as a useful starting point and suggest that quality measurement for each major category will be different. This categorisation is found in the other reviews (Hagen 2000; Ouzounis and Valencia 2003) and mirrors the layers found in typical software systems:



3.2 Categories of tools

3.2.1 Databases and data resources

Searls (2000) identifies databases as being at the heart of genomics, and Anon (2005) noted that it is the aim of databases to allow re-use of data between labs. In terms of technology most have moved from flat files to relational databases, and although there are some that have adopted object-oriented database technology this remains by far the most common platform used.

Searls (2000) describes the lifecycle of web resources. They typically start as an idea by researchers to fill a need and may then follow different paths: they may become disused (though the web site may persist); they may continue to operate and be maintained by the original or like-minded people; they may attract funding for development and maintenance; and they may even become commercialised. These different states may themselves be an indication of quality in terms of how up-to-date or comprehensive the data are, and knowledge of who maintains and funds the database may also be important to potential users of the data: Searls (2000) notes that resources are offered virtually without any warranty other than that gained by inspection and knowledge of the developers' reputation.

There are many hundreds of databases available to the genetics community with a wide range of aims and scopes. A broad categorisation of such tools is: those presenting primary or raw data such as Ensembl¹, Genbank² and DDBJ³; those adding annotation to this data such as Flybase⁴ and SwissProt⁵; and those providing higher level structure and annotation such as Pfam⁶ (Birney, Clamp and Hubbard, 2002). Other types of large-scale data resource used in genetic testing include publication databases such as PubMed⁷ and disease and gene information resources such as OMIM⁸ and Orphanet⁹. Clearly there are overlaps between many of these and they have

1 www.ensembl.org
 2 www.ncbi.nlm.nih.gov
 3 www.ddbj.nig.ac.jp
 4 flybase.bio.indiana.edu
 5 ca.expasy.org/sprot/
 6 www.sanger.ac.uk/Software/Pfam
 7 www.pubmed.gov
 8 www.ncbi.nlm.nih.gov
 9 www.orpha.net

many different aims and scopes. Trying to categorise them to any detailed degree may be difficult and unproductive.

Many databases for genetic testing laboratories aim primarily to collect and share mutation data. Within this more limited scope categorisation may be more meaningful, and is discussed by many authors. All describe core (or central) databases, which deal with mutations for many genes and diseases, and locus-specific databases (LSDBs), which deal with mutations for only one gene or disease, though Porter *et al.* (2000) thought this categorisation an over-simplification. Patrinos and Brookes (2005) also describe National mutation databases, which hold mutation data for a particular population, and other resources such as SNP databases (dbSNP¹⁰, HapMap¹¹, PharmGKB¹²) and PhenomicsDB which holds multi-species phenotype-genotype correlations. Some of the contrasting roles and features of core and LSDBs are summarised in Table 1.

Core (central) databases	LSDBs
More likely to be directly funded and therefore have full-time staff support for informatics and curation, and greater longevity.	Usually an additional project to the researchers' main interest so can become stagnant and may not have funding for expert input e.g. database/informatics support
There are those that validate all submissions, those that present all submitted data, and those that present only published data.	Will contain unpublished and published data – the aim is usually for completeness and high quality.
There are those that present each variation only once and those that present every instance.	Developers and curators are interested in the diseases so are interested in maintaining accurate up-to-date information, presenting a greater depth of knowledge and usually know submitters so are able to capture more data and be responsive to needs
May be infrequently updated, not interactive and have fewer or less flexible facilities such as search tools.	
No central database alone is sufficient for the needs of medical genetics: the approach is for 'mile wide and inch deep' coverage.	Typically aim to support a clinical or diagnostic service: the approach is for 'inch wide and mile deep' coverage.
Interface will be consistent and there is the opportunity to carry out aggregated analyses across data. There may not be the flexibility to represent data relevant to certain loci.	Content and quality extreme variability between genes.
Scale: HGMD (www.hgmd.cf.ac.uk) held 47889 mutations in 1885 genes on 02/09/2005	Typical LSDBs such as HbVar (globin.cse.psu.edu/globin/hbvar/), FXI (www.factorxi.com) and PAX6 (pax6.hgu.mrc.ac.uk) held 1234, 181 and 309 entries respectively on 11/11/2005.

Table 1. Roles and features of core and locus specific databases (LSDBs)

10 www.ncbi.nlm.nih.gov
 11 www.hapmap.org
 12 www.pharmgkb.org

References: Kalmar (2005), Claustres et al (2002), Scriver et al (1999) and (2000) Patrinos and Brookes (2005), Beroud et al (2005).

Claustres et al (2002) and Patrinos and Brookes (2005) noted that LSDBs and core databases have a similar function but are complementary because of their different depths, with each benefiting from the other. Cuticchia (2000) thought curation of central databases in detail is impossible, making it necessary for LSDBs to be established to provide this depth of information for a gene. Brown and McKie (2000) noted the same point: the advantages of a core database are a consistent user interface and comprehensive dataset but the limitations are depth, and that they require considerable computer and human effort to maintain – this is overcome by LSDBs but only for specific genes. Scriver et al (1999) and (2000) produced a set of recommendations for the content, structure and deployment of mutation databases. To be classed as a knowledge base an LSDB must combine scientific and diagnostic data with information useful to clinicians or students, and information for patients and their families. They recommended that LSDBs use HGVS guidelines for content and nomenclature to reduce variability and noted that consortium operated sites have better potential viability and that it is desirable to maintain data security independently of curator's funding. Porter et al (2000) also noted that standardisation is a common issue and that LSDBs should use HUGO gene names¹³, HGVS nomenclature¹⁴ and RefSeq¹⁵ reference sequences.

Cotton and Horaitis (2000) discuss issues of databases and quality control, and treat the issue of data quality separately from that of the tools in question. Quality is equated to accuracy for the data and critical points in the laboratory and interpretation process are identified. Although such laboratory issues are probably outside the scope of this work it does highlight the need for such data to be collected and presented to the user to be able to judge the quality of data, and for validation of data where they have been interpreted e.g. aa residues. They recommend standardisation of the submission and review process for journals, using standardised data submission forms and tools for checking such as Mutation Checker from EBI. Thus the tools and facilities provided by a database or journal will affect the ability to implement quality control. They go on in Horaitis and Cotton (2004) to identify that data should be up-to-date and complete in mutation databases (data quality issues), with ready access and search facilities (tool quality issues). Some of the quality issues identified for mutation data were that nomenclature were non-uniform and central database curators were not expert in the gene. The LSDB generic system that they have developed (LSDB in-a-box) aims to be easy to install, platform independent, handles the core data identified as necessary for good quality and allows extension. Birney, Clamp and Hubbard (2002) are also concerned with the quality of data: some of the issues that they raise include that data are most valuable when systematically organised and integrated; that maintaining evidence trails is important so that derived data are linked to their evidence and changes are propagated appropriately; users need to know whether data are automatically derived, based on experimentation or curation and their accuracy, and there is an need to make measures of accuracy understandable in the context of their use.

Stenson et al (2003) describe the HGMD¹⁶ central database and this highlights some important quality issues regarding the scope and aim of databases and the policies adopted. For example, the types of mutation that this database holds are given, as is the coverage, the way that

13 www.gene.ucl.ac.uk/nomenclature

14 www.hgvs.org

15 www.ncbi.nlm.nih.gov/RefSeq/

16 www.hgmd.cf.ac.uk

different mutations are represented, the policy of only recording each mutation once, the nomenclature standards used, the acceptance policy for data and details of the collaboration with their sponsors and how this affects the access to the data. Cuticchia (2000) also discusses the importance of making access policies and data sources clear along with the mission aims and funding. The importance of continued development and maintenance is stressed as is quality control of the development process and measures to guard against loss of expertise and data – more than one person has knowledge of the resource and it is documented and there are backup and archiving processes in place, and also that the project has long-term viability in terms of funding.

There are many descriptions of mutation databases in the literature: the journals *Bioinformatics* and *Human Mutation* have regular slots for databases. Many of these are for LSDBs or yearly updates from central databases. Recent examples include Beroud et al (2005), Brandon et al (2005), Hamosh et al (2005), Beysen et al (2005), Heinritz et al (2005), Leonard et al (2005), Tzoulaki et al (2005), Saunders et al (2005) and Scriver et al (2003). Some of the common factors that can be drawn from these papers which may be applicable to quality measurement include:

- Data source – published, submitted, collations
- Curation policy
- Provision of search tools
- Extent of linking
- Maintenance policy
- Scope of the data presented – what type of data is it
- How much data is there?
- How fast is it growing?
- When was it last updated?
- How many users are there?
- How does it overlap with other resources?
- What are the limitations? – e.g. OMIM does not have rearrangements
- What standards are used
- What is the funding basis and how long is this secured for
- What are the data security arrangements
- Who are the target users of the database
- What is it for
- Aim or mission
- Implementation – tools, methods etc.
- Nature of interface
- Nomenclature used
- Reference sequence used
- Data display methods – graphical, text
- Categories of mutations presented or how they are labelled.
- Architecture
- QC procedure
- goals

3.2.2 Search and analysis tools

Searls (2000) discusses the range of tools that have been developed – standard tools include alignments, profiles, phylogenetic trees and gene finding, but there are many more. One problem with quality assessment is that there are many different types of tool each with a specific area of application. Even widely used tools such as BLAST have variants which are designed for specific applications, as well as new versions which aim to improve performance,

and even though they may be based on the same algorithm or technique their quality may vary due to differences in programme design and the ability of the user to adjust and understand tool parameters. For these reasons it may be reasonable to suggest that one set of quality measures for such tools is a clear statement of what the tool is for, how it should be applied, what algorithm or approach it uses. Another set of measures can be based around performance measures of the tool – e.g. sensitivity and specificity, speed etc.

The European Bioinformatics Institute (EBI 2005) provides a directory listing a large range of tools and databases. The categories and instances of tools are:

- [Similarity and Homology,](#)
 - [Blast2 - ASD,](#)
 - [Blast2 - EVEC,](#)
 - [Blast2 - NCBI,](#)
 - [Blast2 - Parasite,](#)
 - [Blast2 - WU,](#)
 - [Fasta,](#)
 - [Fasta - ASD,](#)
 - [Fasta - LGIC,](#)
 - [Fasta - Geno./Proteo.,](#)
 - [MPsrch,](#)
- [Prot. Function. Analysis,](#)
 - [CluSTr,](#)
 - [GeneQuiz,](#)
 - [InterProScan,](#)
- [Proteomic Services,](#)
 - [Dasty,](#)
 - [UniProt DAS,](#)
- [Sequence Analysis,](#)
 - [Align,](#)
 - [ClustalW,](#)
 - [GeneWise,](#)
 - [PromoterWise,](#)
- [Structural Analysis,](#)
 - [DALI,](#)
 - [DaliLite,](#)
 - [Maxsprout,](#)
 - [MSD Services,](#)
 - [MSDfold,](#)
- [Tools Miscellaneous,](#)
 - [EMBL Computational Services,](#)
 - [Expression Profiler,](#)
 - [NEWT,](#)
 - [QuickGO,](#)
 - [Readseq,](#)
 - [Web Services,](#)

This shows how there are different versions of the most common tools like BLAST which are aimed at different situations and users, which in turn emphasises the need for users to be provided with information about the aim and use of each tool. EBI provides both a consistent user interface (UI) for the use of each tool and an extensive tool-specific help section, as shown in Figure 1 and Figure 2.

NCBI-Blast2 Protein Database Query

BLAST stands for **B**asic **L**ocal **A**lignment **S**earch **T**ool. The emphasis of this tool is to find regions of sequence similarity, which will yield functional and evolutionary clues about the structure and function of your novel sequence. [WU-BLAST 2.0](#) and NCBI BLAST2 are distinctly different software packages, although they have a common lineage for some portions of their code, so the two packages do their work differently and obtain different results and offer different features. You can also check for vector contamination with [Blast2 EVEC](#).

YOUR EMAIL	SEARCH TITLE	RESULTS	DATABASE	PROGRAM
<input type="text"/>	Sequence	interactive ▾	Protein ▾ UniProt ▾	blastp ▾
ALIGN VIEWS	MATRIX	EXP.THR	FILTER	DROPOFF
pairwise ▾	blosum62 ▾	default ▾	false ▾	default ▾
OPENGAP	EXTENDGAP	GAPALIGN	SCORES	ALIGNMENTS
11 ▾	1 ▾	true ▾	default ▾	default ▾

Enter or Paste a PROTEIN ▾ Sequence in any format: Help

Upload a file: Browse... Run Blast Reset

Figure 1. EBI user interface for NCBI Blast2

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
<input type="text"/>	Sequence	interactive ▾	full ▾	single ▾
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def ▾	def ▾	percent ▾	def ▾	def ▾
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def ▾	def ▾	def ▾	def ▾	def ▾

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers ▾	aligned ▾	none ▾	off ▾	off ▾

Enter or Paste a set of Sequences in any supported format: Help

Upload a file: Browse... Run Reset

Figure 2. EBI user interface for ClustalW

The EBI User Interfaces show several things: that the user interface and the tool are separate, meaning the quality of a tool can depend on which user interface is chosen; that there will always be specific parameters for each tool type; and that users will benefit from tailored help especially in choosing the correct parameters. The help provided by EBI in fact goes into the different BLAST tools and what each is for, allowing the user to choose the right one.

EMBOSS also provide a suite of bioinformatics tools for the molecular biology community (<http://emboss.sourceforge.net/index.html>). These are categorised into groups as shown in Table 2.

Group	Description
Acid	Acid file utilities
Alignment consensus	Merging sequences to make a consensus
Alignment differences	Finding differences between sequences
Alignment dot plots	Dot plot sequence comparisons
Alignment global	Global sequence alignment
Alignment local	Local sequence alignment
Alignment multiple	Multiple sequence alignment
Display	Publication-quality display
Edit	Sequence editing
Enzyme kinetics	Enzyme kinetics calculations
Feature tables	Manipulation and display of sequence annotation
HMM	Hidden Markov Model analysis
Information	Information and general help for users
Menus	Menu interface(s)
Nucleic 2d structure	Nucleic acid secondary structure
Nucleic codon usage	Codon usage analysis
Nucleic composition	Composition of nucleotide sequences
Nucleic CpG islands	CpG island detection and analysis
Nucleic gene finding	Predictions of genes and other genomic features
Nucleic motifs	Nucleic acid motif searches
Nucleic mutation	Nucleic acid sequence mutation
Nucleic primers	Primer prediction
Nucleic profiles	Nucleic acid profile generation and searching
Nucleic repeats	Nucleic acid repeat detection
Nucleic restriction	Restriction enzyme sites in nucleotide sequences
Nucleic RNA folding	RNA folding methods and analysis
Nucleic transcription	Transcription factors, promoters and terminator prediction
Nucleic translation	Translation of nucleotide sequence to protein sequence
Phylogeny consensus	Phylogenetic consensus methods
Phylogeny continuous characters	Phylogenetic continuous character methods
Phylogeny discrete characters	Phylogenetic discrete character methods
Phylogeny distance matrix	Phylogenetic distance matrix methods
Phylogeny gene frequencies	Phylogenetic gene frequency methods
Phylogeny molecular sequence	Phylogenetic tree drawing methods
Phylogeny tree drawing	Phylogenetic molecular sequence methods
Protein 2d structure	Protein secondary structure
Protein 3d structure	Protein tertiary structure
Protein composition	Composition of protein sequences
Protein motifs	Protein motif searches
Protein mutation	Protein sequence mutation
Protein profiles	Protein profile generation and searching
Utils database creation	Database installation
Utils database indexing	Database indexing
Utils misc	Utility tools

Table 2. EMBOSS categorisation of bioinformatics tools

Even at the group level the categories show that these tools are specific to particular analyses.

The need for consistent user interfaces and help that EBI have identified is supported by our experience of employing missense mutation prediction tools – we found difficulties firstly with understanding how to apply the tools, in terms of getting input data in the right format, understanding the options for using the tools and the parameters for their use. We then found it equally difficult to interpret the outputs in terms of understanding whether they represented a correct application of the tool, and there was a lack of understanding of the reliance that could be placed on the results. This issue is addressed in the following paper.

Chavali et al (2005) address the quality of tools in terms of their performance. In their paper they discuss the relative performance of primer design tools, motivated by the apparent variability of T_m predictions by such tools and the ignorance of many users of the limitations of the software that they are using. Their approach was to develop a benchmark test that was applied to each tool with the results analysed statistically. These are presented with the aim of allowing users to employ tools with minimum deviation and to present the limitation and restrictions of tools to the users. In this latter respect the approach is like that of EBI (2005): inform the user to help them use the right tool in the best manner. But it is the provision of statistical information and benchmark testing that makes this paper unlike some others which present new tools or techniques. A brief inspection of some of the many new tools which are presented in journals each month shows that there are not always figures given for their sensitivity or specificity, or performance against benchmark tests. An example is Gu *et al.* (2005) which describes a tool for graphical comparison of haplotype blocks: the details of the motivation for the work, the aims, the language, platform, algorithms, result reporting and UI are given and in these respects this is the same as the information that is required for databases. Ferrer-Costa *et al.* (2005) do provide this information: for their tool for annotation of pathological mutations on proteins they include an assessment of accuracy, the training set used, the archetypes, what the output is and modes of operation. The training set and archetypes are important pieces of information which tell the users how the tool was developed and the gene that was the original target for development.

3.2.3 Interface and visualisation tools

These have developed in response to the large amount of data in genomes (Searls 2000). They include GUIs for data visualisation and interaction, which in some cases have developed into programming interfaces, toolkits and 'widgets'. They are separated from other tools in that they do not carry out analysis of data in themselves but present such data to users. Often however such interfaces will be incorporated into analysis tools and it may be difficult to separate them for quality assessment. Some of the quality issues will indeed be the same as those for other tools so it may be appropriate to deal with both analysis and visualisation tools in a similar way. One quality factor that may be especially relevant to this type of tool though is that of limitations – tools may have limitations to the extent of the data that they can display which should be made clear to the users to avoid the possibility that users assume data do not exist because they are not displayed.

3.3 Approaches to rating

In this section I will propose some different ways that the quality of databases and tools can be assessed.

3.3.1 Feature lists

Many of the papers presented above suggest that one way to allow tools to be assessed is to provide a clear statement of the features of a tool. This could include items such as who has developed the tool, how is it maintained, who funds it etc. Anon (2005) adds the project size and sustained funding in this category. Another set of measures could be based around the data, i.e. what is its provenance (or whether this data is collected and given), how has its quality been assured, how up-to-date and complete is it etc. The approach of EBI is to make this type of information available through the information that they make available about each tool or group of tools. In terms of databases the information could also include the depth of data in a database, e.g. contents may include the mutation alone, population specific allele frequencies, phenotype data. Other quality factors could concern the service or features of a database: tools for data submission, a credit and citation system, review for data validation and a browser for visualisation. Coordination and standardisation are important factors. These features may therefore be based around the features of the tool – standards used, contents, tools and features for submission and visualisation etc.

An analogy for this approach is that of the patient information given with medicines. These typically list, in an easy to understand and standardised format, what the name of the drug is, who makes it, what it is made from, what it is for, what it is NOT for, who should and should not take it, how it should be used etc. This information could be given in a similar way for any tool of database: what data it presents or what it is for, who maintains and funds it, how to use it, what method, algorithm or dataset it is based on, how to interpret the results etc. One important question is whether it should be the task of the tool developer to provide the information or whether it should be gathered by a third party as EBI do, and if the latter approach is taken whether it is practical to develop guidelines on how to use the tool.

A further analogy is given by hotel rating systems: these are based not on how good any hotel is but on how many of the required services or features are provided by each hotel. So, provided a hotel has a certain number of en-suite rooms, restaurant, 24-hour concierge etc. it can expect to receive a certain number of stars. It may be thought attractive to introduce a star rating system for bioinformatics tools: alternatively it may be more meaningful to provide a clear categorisation of tools and a list of the features provided. It is important to avoid the popular misinterpretation of star ratings as a subjective indication of quality.

A more applicable parallel may be drawn with the myGRID project (www.mygrid.org.uk). This is a bioinformatics project which seeks to build services that allow data and tool integration, allowing workflows or pipelines of bioinformatics tools to be assembled. One of the requirements of any tool that is to be used in such a service is that its input requirements and parameters are able to be discovered, and its outputs interpreted. This is referred to as 'metadata' and is discussed in more detail later in this report.

Guttmacher (2001) discussed the difficulty in assessing the quality and accuracy of information given in web-based databases and related the issues to those addressed by ethics standards developed for health information websites. It is equally possible to provide excellent as it is to provide biased or misleading web content: it is suggested that basic questions are asked which include:

- Disclosure – does site disclose mission and ownership/support?
- Confidentiality – does it disclose its policy and is it sufficient to safeguard users
- Timely updating – are updates frequent and their details recorded?

- Expertise – does it list its staff and consultants? Is authorship of information clear? Do these people have appropriate expertise?
- Ethics codes – does it subscribe to any recognised code of conduct such as HON, Hi-Ethics principles or eHealth code of ethics.

These could be the basis of a features list for data. The ethical schemes mentioned both employ a set of principles that those who subscribe must abide by. HON (www.hon.ch) have eight principles which are: authority (the providers are medically trained or it is clear that they are not); complementarity (the info is to support not replace physicians); confidentiality (personal data confidential); attribution (source of data is clear); justifiable (claims supported by clear evidence); transparency of authorship (information is clear and contacts are given); transparency of sponsorship (support is clearly identified); honesty in advertising and editorial policy (use of advertising will be made clear). An EU project (EC 2002) is based on these principles. Hi Ethics (www.hi-ethics.org) has: Privacy policies; enhanced protection for health information; safeguarding consumer privacy, disclosure of ownership and financial sponsorship; identifying advertising and content sponsorship by third parties; quality of health information content; authorship and accountability; disclosure of source and validation for self assessment tools; professionalism; qualifications; transparency of interactions, candour and trustworthiness, disclosure of limitations; mechanism for feedback. Although these guidelines are designed for health web sites many of the principles are applicable to data providers for genetic testing and could be adapted into a list of features.

The objective of a feature list approach is clear – it should aim to provide all the information required for a potential user to choose the right tool or database, to apply it correctly and to interpret and report the results fully and correctly.

3.3.2 Subjective assessment

Under this heading I mean any attempt to present a review of tools from the point of view of a user. This could include both the features and presentation of the tool and the quality of its results. An analogy for this approach is that of the Michelin guide – the star rating used by this guide is based to an extent on the assessment of quality made by inspectors, although there is also an aspect of the features provided. Similar approaches are taken by consumer journals offering reviews of products such as cars, cameras, computers etc., although again it is worth mentioning that these journals usually provide features tables to allow comparison of products features as well as a subjective assessment of the less tangible quality aspects.

This approach would require an independent review system to be developed and a credible organisation assembled to make and publish reviews.

3.3.3 External/Objective assessment

Under this heading I would place approaches which build upon the subjective review by developing standardised statistical or benchmarking tests which will measure such statistics as sensitivity and specificity, standard deviations or errors, applicable ranges and other measures applicable to the specific use of each tool type. It may be the case that such measurements are made for a representative or reference set of data, e.g. certain genes or genomic areas, or that the tool is developed for a certain gene (the archetype). It is especially important that this information is presented alongside any assessment.

This approach is compatible with the other approaches suggested above and I think is essential if bioinformatics tools are to be adopted and reported by medical geneticists. It is likely to be

applicable to analysis tools rather than databases, for which details of their coverage and how up-to-date they are may well serve the same purpose but would not be amenable to statistical analysis.

3.3.4 Medical guidelines approach

Clinical practice guidelines are 'systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances' (Institute of Medicine (IOM), quoted in Agrawal and Shiffman, 2001). The systematic nature of these guidelines has allowed ratings instruments to be developed to counter concerns about the standards of scientific evidence used in their development. Agrawal and Shiffman (2001) state that a guideline's quality should be measured by a "prospective evaluation of its effectiveness in achieving its intended health outcomes", but that this type of evaluation is lacking for most guidelines. Instead, an evaluation of the methodology used to develop the guideline and of the content of the resulting document is used. The three key principles in developing high quality guidelines are identified as: they should be multidisciplinary; they should be based on a systematic review of the literature; they should explicitly link their recommendations to the supporting evidence. The IOM proposed an assessment instrument evaluating eight attributes: clinical applicability/scope, clinical flexibility, reliability/reproducibility and validity are concerned with the substance of the guideline; while clarity, multidisciplinary process, scheduled review and documentation are concerned with the development process. Although this instrument was found too complex to implement it has been the basis of other rating instruments.

Although there has been criticism of the role of guidelines in the practice of medicine, the way that their quality is measured may be a good example of systematic quality measurement. It is not a perfect analogy as bioinformatics tools are not systematically developed to assist in clinical care, and may be relied upon to provide information instead of guiding clinical management, so a requirement for literature review for example may not be applicable. However the approach of separating the substance or outcome of the tool or guideline from its development is in common with the separation of the statistical analysis of a tool's performance from its features that I have discussed above, and some of the attributes for each – scope, reliability and validity of outcomes; and clarity, scheduled review and documentation of the development process – reflect the features and measurements that I have suggested.

3.4 Conclusions

In conclusion to the literature survey, I would suggest that there are two aspects to quality assessment of tools that should be developed.

Firstly, the user needs information in a standardised format about the development process of the tool: what it does, what it is for, who developed it, how it will be supported, the development process used etc. I have found a range of data of this sort in relevant papers but it is unstructured and tends not to categorise the tools with others. The effects on quality of this information are manifold: the correct choice of tool and choice of input parameters is vital to the correct use of the tool, and may only be understood once the purpose and basis of a tool are understood; tools must be used correctly in order to obtain correct results; without information about the provenance of a tool its reliability can only be taken at face value; and the level of service offered to others, such as a diagnostic testing service, can only be as good as the level of service offered by the tools and techniques it uses. EBI provide a good example of categorisation of tools, which allows for the choice of the best technique available for a particular task, and add their own user interfaces and help files to each tool to ensure that the operation and basis of each one is known.

I would suggest that there are three areas that need to be addressed: information about the tool; information about data; and information about operation (i.e. parameter setting). Not all of these will apply to every tool used, e.g. data issues will only apply to databases or tools based on interrogation of a dataset.

The second aspect of quality is the assessment of the performance of a tool, e.g. in terms of sensitivity and specificity, or completeness. This is vital for the correct interpretation of the results, which in turn is vital for the confidence of the user in using and reporting the results.

How can these approaches be applied? The development of quality assessment methods for medical guidelines shows that instruments – systematic checklists based on agreed principles for the development of valid guidelines – can be used to assess the quality of development and I would suggest that a similar approach may be taken for the assessment of development process. This may also be thought of as a ‘feature list’ approach in which features which are agreed to be important for quality are identified for each tool. Information about operation can be addressed through development of information for the user such as that provided by EBI. Assessment of performance requires standardised or benchmark tests – this may require standard reference test sets or scenarios to be developed. Although some of this effort may be achieved by the tool developers it seems likely that an independent body will be required for this work.

4 Survey of testing laboratories

4.1 Method

This survey was undertaken to discover which tools are being used in testing laboratories and how they are used. It was conducted with representatives from a selection of UK testing laboratories then extended to EuroGentest partners. The survey included cytogenetic and molecular genetic laboratories and biochemical services. It consisted of an interview based around the following questions, though other issues were pursued where appropriate:

1. Tool Details
 - a. Why do you use a particular service or tool?
 - b. Who developed it?
 - c. Web based bespoke or proprietary stand alone?
 - d. Is the tool suitable for the task, was it originally designed for this use?
 - e. Do you know how well the tool performs for this task?
 - f. Is it designed for diagnostic testing?
 - g. Are you aware of the alternatives for these tools?
 - h. Is the tool kept up to date?
2. Use of Output
 - a. How are results stored?
 - b. How are the results reported?
 - c. Are results fed into other tools?
 - d. How much confidence do you attach to the results?
3. Tool Management
 - a. How do you ensure that the latest datasets, versions are being used?
 - b. How did you find out about the current tools that you use?
 - c. Do you have any concerns about the tools you are currently using?
 - d. Are there tools that you would like, but are not available?
 - e. Are there any tools that could be developed to be more appropriate for diagnostic testing?

Summaries of the outcomes of these interviews are shown in Appendix A.

4.2 Results

4.2.1 Survey overview

A wide range of software was found to be employed in genetic testing laboratories. It varies from freely available utility tools (often offshoots from academic projects) such as primer design tools and splice-site recognition tools, to commercially licensed sequence analysis and fragment analysis tools. Also reported were in-house developed analysis tools, for a limited number of tasks. Recently emerged management systems that integrate much of the functionality of previous tools with data management facilities are starting to appear.

There also exists a vast array of disease and locus specific databases holding a variety of information useful for the genetics testing community. The most comprehensive list for these appears to be at www.hgmd.org.

In many cases, existing software in a laboratory has come bundled with various machines. For example, ABI or Beckman Coulter sequencing machines include software that perform various sequence analysis tasks, such as base calling and variant detection. Many of these tasks can be achieved independently by products being marketed by independent software houses. Such software appears for trial in laboratories mostly via word of mouth. In cases such as mutation detection, this raises the issue of the quality of the function the software supports. Given the standards expected of testing laboratories, some form of quality assessment of this functionality must be provided and shared to ensure the tool is clinically useful. As an example, the National Genetics Reference Laboratories in Manchester and Wessex in the UK have undertaken detailed assessments of some tools on behalf of UK laboratories to meet this requirement (Patel and Wallace, 2005a and b; White and Potts, 2006).

Other aspects of the software are also important, such as its ease of use, compatibility, ability to be configured to local settings and training. In one case within the mutation detection task, it was reported that "ease of use" was the significant reason a product was employed rather than the improvement in the quality the detection algorithm afforded.

Similar concerns over quality measures exist in regard to the collection of databases that are in use by the genetics testing community. In particular, features such as how current the database is and how well maintained it is, are of interest.

The natural operating structure of most diagnostic laboratories is such that personnel specialise in a specific disease service. Thus it is not normal for laboratories to have a central repository of links and information regarding various software tools and databases. Knowledge of useful tools and databases are mainly kept individually, either as bookmarks or mental notes.

Open and available resources that present such information are not currently available. Several respondents noted that a trustworthy source to validate the quality of the clinically most critical features of the bioinformatic tools is required, especially for the laboratories that do not possess the resources to make such judgements.

4.2.2 Classification of tools discovered

In collecting the list of software and database mentioned in the survey, it is also desirable to list accompanying features, for use in comparison and assessment of quality. Tasks may overlap

with different software products which have varying approaches to managing data processing. For example, Phenosystems offer a product that develops a wider information systems solution than a more specific tool such as SoftGenetics® Mutation Surveyor™, but both contain a mutation detection function.

General aspects covering most software and data management are of relevance, e.g. the nature of the documentation, the support, the maintainability, interoperability (e.g. different formats), the user interface (UI), speed of execution. There are then more specific features related to the class of software, and its supposed functions and capabilities.

Within molecular genetics testing environments we might divide various tasks by noting at what stage of the diagnostic procedure they operate within. For example, we may consider functions of various software for use in the simple divisions of:

- DNA extraction and amplification procedures
- Fragment analysis procedures
- Sequencing procedures
- Reporting procedures

4.2.3 Summary of tools discovered

Class	Purpose	Tools
Primer design	Design of PCR primers	Oligo Oligo Lite Primer3 Webcutter
Fragment analysis	Sizing of PCR products	Genescan> Genotyper Softgenetics GeneMarker ABI GeneMapper Transgenomic Navigator (WAVE) Spreadsheets
Sequence analysis	Analysis of sequence data	ABI GeneMapper Staden Mutation Surveyor SeqScape Chromas CEQ 8800
Databases	Storing, sharing and display of data	NCBI ENSEMBL HGMD (Cardiff) Beta Globin DB OMIM EDDNAL ORPHANET

4.3 Conclusions

Most of the software that has been identified in laboratories is commercial software that is used in conjunction with specific machinery or processes, and databases. Many laboratories are concerned with the operation of their processes so features such as ease of use and compatibility are rated highly, but trustworthy sources of data about the quality of these clinically important tools are not available. Quality of tools can refer to general aspects such as documentation and support but each class of software has a range of specific aspects to address.

From this work we therefore established that the software we need to address are the tools used within laboratories in the analysis of data generated by specific processes. These tools need to be classified and their quality assessed based on general and specific features of the tools.

These findings concur with those of the literature survey: tools can be classified into the broad categories given, their specific features can be described and metrics devised for measuring the quality of their performance in terms of their outputs. What is different is that the tools involved are generally commercial software designed to support machines and which have multiply capabilities, rather than the generally open source, single function software that is associated with the bioinformatics community. We have therefore proceeded with the proposed approach – classification and description of features and capabilities of tools, and determination of suitable performance measures for each class of tools – but based initially on these commercial packages. The method will be applied to databases following this initial work.

5 Archive development

In the section I will describe how a system was chosen for the archive of tools. From discussion with EuroGentest partners we established that sequence analysis tools should be our initial priority for quality assessment so we have used these tools as the pilot case for implementation. It also became clear that knowledge about each class of tool will be held by specific individuals or groups, so that the initial proposal of forming an expert committee to develop quality measures for all tool types was not feasible: instead, ad-hoc groups with expertise in each tool type should be formed.

5.1 Requirements

The aim of this work is to create an online repository of bioinformatics tools, with the following functions:

- To provide a list of tools
- To categorise and annotate these tools to allow their function and merits to be compared
- To present objective measurements of their performance

The archive will need to be accessible to the genetic testing community, and will need to allow contributions at least from the groups interested in each tool type.

5.2 Design

A web-based tool is an obvious choice given the requirements for accessibility, and there is no need for this to be other than a browser-based application as the level of interaction required – editing of documents, posting of messages etc., are within the scope of such applications. As

there are data to be held, a database backed system is needed, but the amount of data will be relatively low so there is no overriding need for a separate data server.

One of the functions of the system is to allow joint contributions to shared data so there are concurrency issues that will need to be allowed for, e.g. to prevent the same document being edited by two people at once. Security will be needed in order to limit editing of data to those authorised to do so, and access to the data may also need to be constrained.

5.3 Implementation

In considering implementation we looked at two areas: metadata and the system used to present the data. For data and tools to be indexed and annotated it is necessary that information is held about them, and this is referred to as metadata. E-science projects like myGRID (Stevens *et al.* 2003) are developing generalised approaches to this issue in order to allow the function of bioinformatics tools to be detected automatically by other software and thus be used in pipelines or workflows. In order that the metadata used has a shared and standardised meaning they are organised into hierarchical lists or ontologies. Using existing ontologies for this work would be sensible because it avoids unnecessary work and links this project to the broader approaches being developed by myGRID and other such projects. However there appear to be few resources available to address the needs that have been discovered. myGRID is developing the technology to be able to apply metadata to tools, e.g. the Grimoire project, but does not have a detailed classification of tools. The Dublin Core metadata initiative (dublincore.org) is a more generally applicable metadata set used for web-based resources, and this comprises the following elements: title, subject, description, type, source, relation, coverage, creator, publisher, rights, date, format, identifier and language. These can be applied to any web resource and will be a useful standard for the present work, but are not sufficiently detailed for the demands of the project. In addition, they are designed for web resources rather than software tools so will not be universally applicable to the resources being analysed in this work. We have therefore developed our own categorisation of tools for use in this project. At present the list is relatively short and is being used a simple taxonomy of tools: no attempt has been made to develop the logical descriptions that are used in more complex fields such as medical ontologies which allow them to be automatically classified.

The categorisation is in two parts. The first lists all of the tool types discovered in the survey so far: this is shown in Figure 3. The second part provides details about the features that are to be used to annotate the tools description. The latter is being developed as the pilot data are being entered into the prototype system so is less complete: as more tools types are analysed common features or groups of features should become more prominent. Figure 4 shows the feature list currently being developed for Sequence Analysis tools which are the first tool type being studied.

1. **Databases**

These are data repositories that are generally available via the internet.

 1. **Genetic information databases**
 1. **Mutation databases**
 1. **Core databases**

Databases holding mutation data for a range of genes
 2. **Locus specific databases**

Databases holding mutation data for a single or related group of genes
 1. **Knowledge bases**
 2. **Literature databases**
 3. **Medical databases**
 1. **Inherited disease databases**
2. **Tools**
 1. **Design tools**
 1. **Primer design tools**
 2. **Analysis tools**
 1. **Fragment analysis tools**

Tools for size analysis of PCR products
 2. **Sequence analysis tools**

Tools for analysis of sequence data
 3. **Search tools**
 4. **Visualisation tools**

Figure 3. Classification of tool types

We have considered different ways to implement the archive and have chosen to develop a pilot application using a wiki. Wikis are web sites which allow pages of data to be created and edited via a web browser by the users (see <http://en.wikipedia.org/wiki/Wiki>) and thus facilitate collaborative document development. The pages are stored in a database and there are mechanism such as audit trails and the ability to 'roll-back' changes to guard against loss or corruption of data. Wikipedia (www.wikipedia.org) is an example of a project based on a wiki – it is an encyclopaedia in which the users are able to create and edit entries via the pages of the web site itself. Mediawiki is the application used for Wikipedia and has been used in this project. A site has been established at ngri.man.ac.uk/mediawiki. This approach was chosen because it allows a complex web site to be created easily and to be maintained continually, it will allow a group of users to collaborate on sets of data, it is very flexible, allowing any number of web pages to be created to represent any type of information, security settings can be made to limit access and edit rights, and each page has a discussion page that will allow users of the data to add comments and discuss data without affecting the pages. The drawbacks are the lack of rigidity in applying classifications and the knowledge required to edit pages: straightforward text pages are relatively easy to edit, though some learning is required, but where tables have been implemented, editing will be harder and unlikely to be undertaken by anyone who has not become familiar with the wiki. This suggests that a curator should undertake changes of this nature, leaving others to comment on pages or make more straightforward changes to documents.

Feature	Sub-feature	ABI Genemapper	Staden	Mutation Surveyor	SeqScape	Chromas	SEQ8800
Developer				SoftGenetics®			
Maintained by				SoftGenetics®			
Version				2.51			
First version date							
Last updated							
Funding basis				Commercial product			
Analysis algorithm				Anti-correlation			
Operating system				Windows			
Documentation				Manuals, help button			
Support							
Capacity				400, 48 and 24 lane			
Throughput				1 billion base pairs/day*			
Input filetype				scf, ab1, abi			
Modes	Uni-directional			Y			
	Bi-directional			Y			
Detection	Heterozygotes			Y			
	Homozygotes			Y			
Mutation types	Indel			1-100bp de-convolution			
	Low-level mosaic			Manual			
Output files				Text, Excel, XML, HTML			
Mutation nomenclature				HGVS			
Chemistry	Terminator			Y			
	Primer			Y			
Mutant allele sensitivity	Unidirectional			10% of primary peak*			
	Bidirectional			5% of primary peak*			
Claimed Sensitivity	Unidirectional			>=95%*			
	Bidirectional			>99%*			
Measured sensitivity	Long dataset			84.8% (bidirectional)			
	Short dataset			89.9% (bidirectional)			
Positive predictive value	Long dataset			52.2% (bidirectional)			
	Short dataset			60.1% (bidirectional)			

*Claimed

Figure 4. Features list for Sequence Analysis Tools

5.4 Analysis of Sequence Analysis Tools

As mentioned above, we have considered Sequence Analysis tools as a pilot case for developing the quality system. With reference to the conclusions from the literature search, the approach to quality assessment uses the tool categorisation (Figure 3) to show what the aim of the tool is, who it is for etc. in a general way, then the features table (Figure 4) gives more specific information and also presents an assessment of quality, in this case sensitivity data. The features list was developed by inspection of the software and of the details published by the software developers, and by an independent assessment of quality that is discussed below. The features are arranged into groups: firstly there is information about the tool’s development – its developer, version, funding basis, algorithm, platform and support issues like documentation and support arrangements; then there are details of the scope of the tool – capacity, modes, input files, types of mutation detectable, output file types, mutation nomenclature used, sequencing chemistry accepted; and finally there is the quality assessment: sensitivities claimed for mutant alleles, claimed sensitivity for mutation detection, measured sensitivity for mutation detection and measure false positive rates. Each row and column heading are links to web pages that can be used to provide further details.

The aspect of quality that has not been addressed in detail here is the presentation of parameter setting for the tools, i.e. help with correct usage of tools to achieve the best results. Guidance and discussion can be provided using wiki pages created for each tool. Figure 5 shows how a

page is easily created for a tool and can have further pages added for discussion of parameter setting and general discussion of the tool.



Figure 5. Pages can be developed for each tool, with links to pages for help with parameter setting and to a discussion area.

The Technology Assessment for Mutation Surveyor™ undertaken at NGRl(Manchester) in 2005 (Patel and Wallace, 2005b) has been used as the basis for the quality assessment for this tool type. The authors of this report presented the features of the tool and the results of practical tests which used four data sets chosen to represent a range of sequencing platforms, laboratories, chemistries and read lengths. They identified the avoidance of false negatives as the most important requirement for this type of tool, with false positives able to be tolerated if not excessive. False negatives can be expressed by the sensitivity, which is given as the number of true positives divided by the number of true positives plus false negatives, i.e. the correctly identified mutations as a percentage of the actual number of mutations present. The higher the sensitivity, the better able the programme is to detect mutations correctly. Sensitivity has therefore been chosen as the main quality assessment parameter for the feature table shown in Figure 4, with figures entered for Mutation Surveyor from the data presented by Patel and Wallace (2005b). These have been presented for the long read-length data set and as an average for the three short read-length data sets used.

False positives are expressed by the specificity, which is given as the number of true negatives divided by the number of true negatives plus false positives, i.e. the correctly identified normal bases as a percentage of the actual number of normal bases present. Specificity is more difficult to measure in practical tests because of the large number of bases that are scanned and the effect of trace quality on the performance, as not all bases are of high enough quality to be analysed. Specificity has not therefore been used in the feature table, but false positive rates can be high and therefore could be a factor in tool choice. The positive predictive value has therefore been used in the feature table. This is defined as the number of true positives divided by the number of true positives plus false positives, i.e. it is the percentage of positive detections that are correct. Thus a high false positive rate leads to a low positive predictive value.

Having developed the feature table and quality assessment measures for Sequence Analysis tools and entered data for Mutation Surveyor™ it is obviously necessary to complete similar details for other Sequence Analysis tools. For results to be directly comparable with other tools the same data sets should be used to determine the quality assessment statistics, and the NGRl datasets are available for use by others for this purpose. This may not always be possible however as software may be tied to a particular platform. For example, NGRl have also carried out a Technology Assessment on the Applied Biosystems VariantSeq and SeqScape system which is clearly linked to the Applied Biosystems platform (Patel and Wallace, 2005a). It did use one of the datasets used in the Mutation Surveyor assessment though. The other important point to note is that Figure 4 represents only the starting point of the development of this table: it is now necessary to seek input from a group of interested partners to refine the approach until it meets the requirements of the genetic testing community.

6 Conclusion

We have found that genetic testing laboratories are generally using commercial software products linked to analysis instruments and web-based databases. Our literature survey suggested that the quality of tools in bioinformatics can be addressed by firstly classifying the tools to allow assessment of their function, detailing their features to allow a more specific assessment of their attributes, and presenting test or benchmark data to give a subjective assessment of their performance in a diagnostic setting. We have developed an interactive web site which contains a categorical list of software tools and allows tabulated data about each class of tool to be developed. Feedback has suggested that Sequence Analysis tools should be assessed first so a pilot case has been undertaken using the SoftGenetics® tool Mutation Surveyor™, based on a Technology Assessment carried out at NGRL(Manchester). This approach must now be presented to a group of EuroGentest partners with interest and expertise in this area to allow the method to be reviewed and modified if necessary and to develop an approach to assessing other tools in the same class. It must also be extended to other classes of tools, with genetic databases a priority.

7 Bibliography

- Agrawal, A. and Shiffman, R.N. (2001). Evaluation of guideline quality using GEM_Q. *Medinfo* 10(Pt 2):1097-101
- Anon (2005). Editorial: WayStation to HUGOBase. *Nature Genetics* 37(8):783.
- Beroud, C, Hamround, D., Collod-Beroud, G., Boileau, C., Soussi, T and Claustres M. (2005). UMD (Universal Mutation Database): 2005 update. *Human Mutation* 26(3):184-191.
- Beysen D et al (2005). The human FOXL2 mutation database. *Human mutation* 24:189-193.
- Birney, E. et al. (2004) Ensembl 2004. *Nucleic Acids Research* 32: D468-D470.
- Birney, E., Clamp, M. and Hubbard, T. (2002). Databases and tools for browsing genomes. *Annual Review of Genomics and Human Genetics* 3:293-310.
- Brandon M.C. et al (2005). MITOMAP: a human mitochondrial genome database – 2004 update. *Nucleic Acids Research* 33:D611-D613.
- Brown A.F. and McKie M.M. (2000). MuStar and other software for locus-specific mutation databases. *Human Mutation* 15:76-85.
- Chavali S et al (2005). Oligonucleotide properties determination and primer designing: a critical examination of predictions. *Bioinformatics* 21(20): 3918-3925.
- Claustres et al (2002). Time for a Unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Research* 12(5):680-688.
- Cotton, R.G.H., Horaitis, O. (2000). Quality control in the discovery, reporting and recording of genomic variation. *Human Mutation* 15:16-21.
- Cuticchia A.J. (2000). Future vision of the GDB Human Genome Database. *Human Mutation* 15:62-67.
- EBI (2005). EBI Services. www.ebi.ac.uk/services and linked pages.
- EC (2002) eEurope 2002: Quality Criteria for Health Related websites. Brussels, 29/22/2002 Com (2002) 667 final.
- Ferrer-Costa et al (2005). PMUT: a web based tool for the annotation of pathological mutations on proteins. *Bioinformatics* 21(14): 3176-3178.
- Gu S., Pakstis, A J, Kidd, K K. (2005). HAPLOT: a graphical comparison of haplotype blocks, tagSNP sets and SNP variation for multiple populations. *Bioinformatics* 21(20):3938-3939.
- Guttmacher, A.E. (2001). Human Genetics on the Web. *Annual Review of Genomics and Human Genetics* 2:213-233.
- Hagen, J.B. (2000). The origins of bioinformatics. *Nature Review Genetics* 1:231-236.
- Hamosh A et al (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* 33:D514-D517.

Health on the Net Foundation. www.hon.ch.

Heinritz H et al (2005). The human TBX5 gene mutation database. *Human mutation Database in brief #846 Online*.

Hi Ethics. www.hi-ethics.org.

Hoffman M, Arnoldi, C. and Chuang, I (2005). The clinical bioinformatics ontology: a curated semantic network utilizing refseq information. *Pacific Symposium on Biocomputing 2005*:139-150.

Horaitis, O and Cotton, R.G.H. (2004). The challenge of documenting mutation across the genome: the Human Genome Variation Society approach. *Human Mutation* 23:447-452.

Kalmar L. et al (2005). HAEdb: a novel interactive, locus-specific mutation database for the C1 inhibitor gene. *Human mutation* 25:1-5.

Leonard H et al (2005). Genotype and early development in Rett syndrome: the value of international data. *Brain and development (in press)*.

Oetting WS and Tabone T (2005). The 2004 Human Genome Variation Society Scientific Meeting. *Human mutation* 26(2):160-163

Ouzounis, C.A. and Valencia, A. (2003). Early bioinformatics: the birth of a discipline – a personal view. *Bioinformatics* 19(17):2176-2190.

Patel, Y. and Wallace, A. (2005a). Applied Biosystems VariantSEQr™ and SeqScape® v2.1 – an assessment using a model system. *Technology Assessment Report, NGRL (Manchester), January 2005*. www.ngri.org.uk/Manchester/Technologypubs.htm

Patel, Y. and Wallace, A. (2005b). Automated mutation detection using SoftGenetics® Mutation Surveyor™ v2.51. *Technology Assessment Report, NGRL (Manchester), September 2005*. www.ngri.org.uk/Manchester/Technologypubs.htm

Patrinos G.P. and Brookes, A.J. (2005). DNA, diseases and databases: disastrously deficient. *Trends in genetics* 21(6): 333-338.

Porter, C.J., Talbot, C.C. Jr., Cuticchia, A.J. (2000). Central Mutation Databases – A review. *Human Mutation* 15:36-44.

Saunders R E et al (2005). Factor XI deficiency database: an interactive web database of mutations, phenotypes and structural analysis tools. *Human mutation*: 26:192-198.

Scriver CR et al (2003). PAHdb 2003: what a locus-specific knowledgebase can do. *Human mutation* 21:333-344.

Scriver CR, Nowacki PM and Lehvaslaiho H and the working group (2000). Guidelines and recommendations for content, structure and deployment of mutation databases: II. Journey. *Human Mutation* 13:344-350.

Scriver CR, Nowacki PM and Lehvaslaiho H (1999). Guidelines and recommendations for content, structure and deployment of mutation databases. *Human Mutation* 13:344-350.

- Searls D.B. (2000). Bioinformatics tools for whole genomes. *Annual Review of Genomics and Human Genetics* 1:251-279.
- Splendore A et al (2005). *TCOF1* mutation database: novel mutation in the alternatively spliced exon 6A and update in mutation nomenclature. *Human mutation* 25:429-434.
- Staats, B. et al (2005). Genewindow: an interactive tools for the visulisation of genomic variation. *Nature Genetics* 37(2):109-110.
- Stein, L.D. (2003). Integrating biological databases. *Nature Reviews Genetics* 4:337-345.
- Stenson P.D. et al (2003). Human Gene Mutation Database (HGMD): 2003 Update. *Human Mutation* 21:577-581.
- Stevens, R, Robinson, A and Goble, C.A. *myGrid: Personalised Bioinformatics on the Information Grid* in proceedings of 11th International Conference on Intelligent Systems in Molecular Biology, 29th June–3rd July 2003, Brisbane, Australia, published Bioinformatics Vol. 19 Suppl. 1 2003, i302-i304.
- Syvänen, A-C, Taylor, G.R.T. (2004). Approaches for analysing human mutations and nucleotide sequence variation: a report from the seventh international mutation detections meeting, 2003. *Human Mutation* 23:401-405.
- Tzoulaki I, White IMS and Hanson IM (2005). *PAX6* mutations: genotype-phenotype correlations. *BMC Genetics* 6:27.
- White, H. and Potts, G. (2006). Mutation scanning by high resolution melt analysis. Evaluation of Rotor-Gene™ 6000 (Corbett Life Science), HR-1™ and 384 well LightScanner™ (Idaho Technology). *Technology Assessment Report, NGRL (Wessex), June 2006.*
<http://www.ngrl.org.uk/Wessex/downloads.htm>

8 Appendix A

8.1 Survey responses

Executive Summary for each laboratory interviewed

Biochemical Lab

Perform a range of chemical tests of which there are 1000s (see <http://www.metbio.net/>). Some of these tests are performed using instruments and some manually. Data is normally downloaded onto a PC and analysed using ad-hoc software. Pathology systems are often involved in the data flow. The types of software employed are instrument based, e.g TANDEM Mass Spectrometers, quality control software, and image analysis software. Many databases are used, such as Cystic Fibrosis databases, Parent Support group databases.

Within the biochemical testing environment the variety of software and data resources utilised is extremely broad when compared to the more focussed molecular genetics testing laboratories.

Cytogenetics Lab

Their patient system was built to their specifications by a software house. They tend to use imaging software and analysis software that is tightly integrated with any instruments that have been purchased. The responsibility for the quality of the product is seen to reside with the manufacturer, and guaranteed by the licenses and service contracts that are part of the purchased product. All quality testing of outputs are left to the manufacturer. For new versions of products, training is always given.

Health Informatics Developer

There is a common mentality that software tools should be treated similarly to hardware in the genetics testing arena, for example purchasing then re-purchasing software. Software tools and databases are certainly not interoperable. For this reason standards should be of strong interest to the genetics testing community. New software tools penetrate the laboratory environment largely through word of mouth.

Genetics Testing Lab 1

Utilise the primer design software Primer 3, and general sequence databases such as Ensembl and NCBI. Their sequencing machines came with the Beckman Coulter sequencing software, which they were happy with. However they are looking to move towards a new software system, and are involved in beta testing Phenosystems products. They use and contribute to a variety of public mutation databases, e.g HAMSTeRS <http://europium.csc.mrc.ac.uk/WebPages/Main/main.htm>, DBD <http://www.dmd.nl/main.html>)

Genetics Testing Lab 2

Software is fragmented into groups that deal with specific diseases. A number of trials for various mutation detection software tools such as Seqscape are under way, but nothing has been finalised. Tool management and quality is mostly left up to the users of the software.

Genetics Testing Lab 3

For primer design use web cutter. Use the ABI software, Gene scan, Gene mapper, Gene marker. Used Mutation Surveyor for ease of use and ease of installation. SeqScape had problems with its installation, so it was not continued with. In-house software has been created for small things, such as di-nucleotide repeats, fragment size and calculating the quantity of DNA in an extraction process.

In general, useful software and web resources are managed using bookmarks on a browser. Their focus for software is to create a pipeline that automates tasks from the instrument directly to reporting.

Genetics Testing Lab 4

Standard set of primer tools design tools are used (Oligo and Primer3). Software for sequencing composed of the default software that arrived with the sequencing machines (ABI). Mutation Surveyor was employed, mainly on the basis of word of mouth. There are some analysis programs that have been created in-house.

There is no systematic organised repository of tools or references. There is also no systematic procedure in testing and checking the output for such tools, likewise for seeking updates for such tools. From a brief audit of the computers in their lab, there appears to be a lot of software downloaded by individuals, either tried or used once (such as demonstrations) but then discarded or forgotten about.

Genetics Testing Lab 5

Makes the point that they, and many others, have very little in the way of software support for their procedures. The problem with employing these tools is that they need to understand the quality and limitations. Hence they have delayed on purchasing software such as Mutation Surveyor. The participant emphasised that the feature of language support is very important in European genetics testing environments.

Genetics Testing Lab 6

Use standard primer design tools, for sequencing they use SeqA, Seqscape, Chromas and general sequence analysis tools. They have no central repository for storing links to Databases. They are generally happy with their procedures and the software that they employ.

Genetics Testing Lab 7

Standard primer design software are used (Oligo and Primer 3) as well as tools to analyse repeats and overlaps with SNPs. Tools to aid fragment analysis are Genescan / Genotyper. Software for pedigree analysis such as Mlink are also employed. Sequence tools used are Staden and MutationSurveyor, however these do not allow for automated detection of mutations. Further tools may be used when checking whether the mutation might have an effect on the expressed gene. They use Alex Dong's Splice site, Gene Splicer, Spliceview, other general sequencing tools such as multiple alignment to look for conserved residues that may indicate that a mutation resides on an important residue. Being able to name and submit mutations is inefficient and needs to be automated.

Genetics Testing Lab 8

They use GeneScan / Genotyper for fragment analysis. When using restriction fragments, they check for other known mutations. They are happy with the ABI machines and software that are presently employed, and no alternatives are seen to be necessary. They use "linkage" software suite to allow for a Bayesian calculation of risk for a disease.