

ARTICLE

# A new web-based data mining tool for the identification of candidate genes for human genetic disorders

Marc A van Driel<sup>\*,1,2</sup>, Koen Cuelenaere<sup>1,3</sup>, Patrick PCW Kemmeren<sup>1,4</sup>, Jack AM Leunissen<sup>1,3</sup> and Han G Brunner<sup>2</sup>

<sup>1</sup>Centre for Molecular and Biomolecular Informatics, University of Nijmegen, Nijmegen, The Netherlands;

<sup>2</sup>Department of Human Genetics, University Medical Centre Nijmegen, Nijmegen, The Netherlands; <sup>3</sup>Dalicon BV, Nijmegen, The Netherlands; <sup>4</sup>Genomics Laboratory, University Medical Centre Utrecht, Utrecht, The Netherlands

To identify the gene underlying a human genetic disorder can be difficult and time-consuming. Typically, positional data delimit a chromosomal region that contains between 20 and 200 genes. The choice then lies between sequencing large numbers of genes, or setting priorities by combining positional data with available expression and phenotype data, contained in different internet databases. This process of examining positional candidates for possible functional clues may be performed in many different ways, depending on the investigator's knowledge and experience. Here, we report on a new tool called the GeneSeeker, which gathers and combines positional data and expression/phenotypic data in an automated way from nine different web-based databases. This results in a quick overview of interesting candidate genes in the region of interest. The GeneSeeker system is built in a modular fashion allowing for easy addition or removal of databases if required. Databases are searched directly through the web, which obviates the need for data warehousing. In order to evaluate the GeneSeeker tool, we analysed syndromes with known genesis. For each of 10 syndromes the GeneSeeker programme generated a shortlist that contained a significantly reduced number of candidate genes from the critical region, yet still contained the causative gene. On average, a list of 163 genes based on position alone was reduced to a more manageable list of 22 genes based on position and expression or phenotype information. We are currently expanding the tool by adding other databases. The GeneSeeker is available via the web-interface (<http://www.cmbi.kun.nl/GeneSeeker/>).

*European Journal of Human Genetics* (2003) 11, 57–63. doi:10.1038/sj.ejhg.5200918

**Keywords:** data mining; candidate gene prediction; bioinformatics

## Introduction

Positional cloning and candidate gene analysis are commonly used as complementary strategies for the identification of genes involved in human genetic disorders. With the recent completion of the human genome draft sequence a comprehensive list of positional candidate genes can often be obtained. For many diseases the critical inter-

val will be between 0.5 and 10 cM, with the number of genes anywhere between 5 and 300. Prioritising these genes for mutation analysis is the logical next step. This requires that the researcher collects information from various sources on expression patterns, biological function, animal models, related human diseases and other relevant data. Clearly, researchers differ widely in their ability to retrieve relevant information that is stored in a growing number of separate (and often unlinked) on-line databases. Moreover, this process tends to be very time-consuming, and many hours may go into collecting and sorting the relevant information. Integrating information from the databases in

\*Correspondence: MA van Driel, Centre for Molecular and Biomolecular Informatics (CMBI), Faculty of Science, University of Nijmegen, P.O. Box 9010, 6500GL Nijmegen, The Netherlands. Tel: +31 243653391; Fax: +31 243652977; E-mail: M.vanDriel@cmbi.kun.nl

Received 1 July 2002; revised 1 October 2002; accepted 9 October 2002

an automatic way would allow researchers to get a quick snapshot overview of their particular candidate region.

Here we report on a new bioinformatics tool, which gathers both positional as well as expression/phenotypic data in an automated way from nine different databases and then combines this information using Boolean operators. This results in a quick overview of candidate genes in the genetic region of interest. The GeneSeeker system is built in a modular fashion, making it easy to maintain and expand. A further advantage is that there is no need for data warehousing or updating because the databases are searched directly through the web.

In its present form, the GeneSeeker tool uses the Genome Database (GDB)<sup>1</sup> and the Online Mendelian Inheritance in Man (OMIM (URL: <http://www.ncbi.nlm.nih.gov/omim/>)) to obtain human mapping data. Genetic localisations specified by the user are also translated with the aid of an 'Oxford-grid', to search the appropriate mouse databases (eg the Mouse Genome Database (MGD))<sup>2</sup>. The key tissues affected by the genetic disorder are used to query phenotypic or expression related databases, including the OMIM phenotype fields, Swissprot,<sup>3</sup> and Medline (National Library of Medicine, Bethesda, USA) for data on human phenotypes and the Gene Expression Database (GXD),<sup>4</sup> the Transgenic/Targeted Mutation Database (TBASE),<sup>5</sup> and the Mouse Locus Catalog (MLC)<sup>2</sup> for gene expression patterns and phenotypes in mice. A general overview of the data flow within the programme is given in Figure 1.

## Materials and methods

### The GeneSeeker interface

The homepage of the GeneSeeker (<http://www.cmbi.kun.nl/GeneSeeker/>) allows the user to specify the genetic mapping information. This can be a chromosome, a chromosome arm, or a range (eg 7p15-7p14). If necessary, a combination of genetic localisations can be entered (eg 4p16-4p14 or 4q31-4q35). Gene expression or phenotypic information can be entered in a separate box, in which the user specifies the tissue names where either direct RNA expression or phenotypic expression of the candidate gene is expected. For example, the phenotype of Hand-Foot-Uterus Syndrome<sup>6</sup> can be translated into the expression terms 'limb or genital'. Advanced options include a thesaurus,<sup>7</sup> which can be used to include alternatives and hence broaden the expression search term. In case of 'limb or genital' use of the thesaurus will result in 'upper limb' or arm or limb or joint or 'lower limb' or 'hip joint' or toes or digit or 'male genital' or testes or testis or 'Sertoli cells' or 'female genital' or ovaries or ovary or uterus or vagina'.

A number of refinement options have been implemented, such as the possibility to exclude databases, to exclude housekeeping or user-specified genes, to change the maximum distance for the Oxford-grid (used in Human-to-Mouse map translation, see below), and to put multiple searches in a batch list.

### Databases used

The GeneSeeker searches three types of databases: genetic localisation, gene expression, and phenotypic databases. GDB and MIMMAP (a reformatted version of the OMIM gene mapping information) are searched for genes in a specified chromosome location in humans, while MGD is queried for mouse genes in the homologous regions. From the gene expression and phenotypic databases GXD, SWISS-PROT, TrEMBL, MLC, OMIM, TBASE and Medline, all the genes are extracted which match the given expression terms. Database web addresses are given in Table 1.

### Gene naming

Different databases cause the data retrieved to be in different output formats. In contrast, communication between programme processes, logical combination and analysis of the data obtained require a uniform nomenclature. To circumvent this problem a list of synonyms was created using the gene-name information stored in SWISSPROT in combination with GDB's 'alias' information. This synonym list is updated weekly, and its use should remove a number of potential naming problems. As an exception, the gene-naming process in TBASE is highly variable and this could not always be neutralised by the use of these lists.

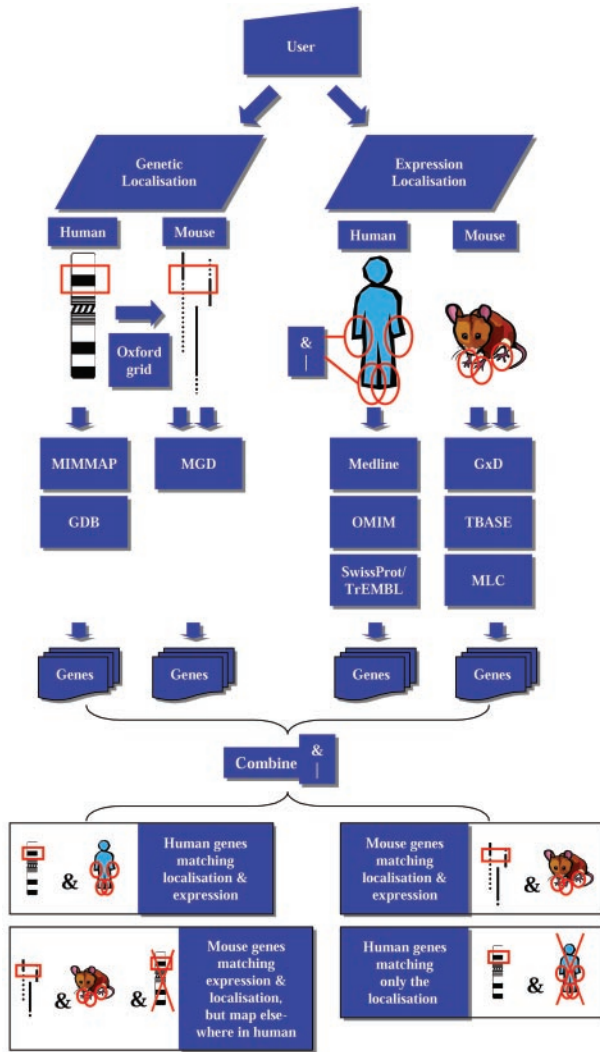
### Extraction of genes from the databases

Human gene names are selected upon the fact that they must consist of one or more capital letters and/or numbers in the databases. Mouse genes begin with a capital letter, followed by one or more lowercase letters or numbers. The id-numbers of the genes in the databases are taken as a unique identifier. In cases where mouse databases are used (TBASE/MGD/GXD), the mouse gene names are translated into human gene names, using a list obtained from GXD. The obtained gene names are compared with the synonym list, obtained from SWISSPROT. If a synonym exists, this then replaces the gene name. The gene names are reported to the GeneSeeker programme together with an URL-encoded link (Uniform Resource Locator) to the entry.

### Databases containing locus information

The MIMMAP and GDB databases are searched for all genes between two human genetic locations, including any overlapping genes.

In order to be able to search mouse gene localisation databases, human genetic localisations are converted to mouse localisations by an 'Oxford-grid' as supplied by the MGD.<sup>2</sup> Mouse genes are searched positionally rather than on similarity, since the goal is to find extra genes. The input for the 'Oxford-grid' is a human chromosome number with a band range (eg 7p15-p21). This location is then translated into mouse chromosomes with a chromosomal range in cM. Two locations are taken as one range if they are closer to each other than specified in the maximum distance. If not, they are returned as two separate regions. Each region



**Figure 1** A general overview of the GeneSeeker programme. The query entered by the user is pre-processed for Human and Mouse databases and subsequently reformulated into the format appropriate for each database. The database queries result in lists of genes, which are combined by Boolean operators according to the query as formulated by the user. The results are presented in the four boxes at the foot of the figure.

is returned with a standard extension of 5 cM, so as not to miss any genes located on the region boundaries. The output from the 'Oxford-grid' contains mouse chromosomes with their cM-range and is reported back to the GeneSeeker main programme. This range is subsequently used to query MGD for homologous mouse genes.

**Databases containing gene expression or phenotypic information**

For all genes expressed in a certain tissue type or associated with a phenotypic feature involving a specified tissue or organ the description and comment lines, matching

'human' are extracted from the SWISSPROT, SWISSNEW, TrEMBL, and TrEMBLNEW databases. The gene names are selected from the gene-name field.

The same procedure is followed for the Medline database, using an advanced Boolean search for all genes expressed in a certain tissue type or sharing a phenotypic feature of a syndrome in humans (query: 'tissue *and* human[orgn] *not* mouse[orgn] *not* rat[orgn]'). The human gene names are selected by the fact that these begin with two capital letters, followed by one or more capital letters or numbers. Common abbreviations such as DNA, RNA, PCR, and others are filtered out.

All genes expressed in a certain tissue type or with a phenotypic feature of a syndrome are extracted from the TBASE database using the 'phenotype' field and selecting the mouse as the organism, from GxD and OMIM using the 'abstract' field and 'text' field respectively, and from the MLC database using the 'phenotype' field. The obtained mouse gene names are translated into human gene names, using a list obtained from MLC itself.

**Test selection of human genetic disorders**

To test the ability of the GeneSeeker programme to identify candidate genes, eight syndromes with known genesis where used. We also evaluated two syndromes whose genesis at the time of the query had not yet been published: Acro-Dermato-Ungual-Lacrimal-Tooth (ADULT) syndrome<sup>8</sup> and Noonan syndrome<sup>9,10</sup> (Table 2). These two examples presented an excellent opportunity to test the system without the noise from direct pointers to the gene in the databases used.

**Querying the GeneSeeker**

The setup of the GeneSeeker makes it possible to submit queries in a number of ways. To benchmark the performance, accuracy, and the flexibility of the system, the same query was formulated in different ways. Each syndrome mentioned in Table 2 was queried using primary expression terms (Table 2) combined with the Boolean operators *and* or *or* for all terms. For example Alagille syndrome was formulated once as 'liver *and* eye *and* heart', and also as 'liver *or* eye *or* heart'. In addition, in some queries the thesaurus/embryological terms were used. Thus, 'eye' became (eyes *or* eye *or* conjunctiva *or* cornea *or* lens *or* optic nerve *or* retina *or* vitreous *or* 'conjunctival sac').

**Evaluation**

Each result is saved as a HTML file in a separate directory, containing the output from the different databases analysed by the GeneSeeker. The output of the analysis is presented in four tables. (1) A list of human genes in the correct genetic region and matching the specified expression profile, (2) a list of mouse genes matching the syntenic region(s) as well as the expression profile, but with no matching human gene name, (3) a list of mouse genes

**Table 1** Database URL's. The number of entries is based on the query formulation used by the GeneSeeker to extract human/mouse related information, and thus can differ from the total number of entries in the database

<i>Data bank</i>	<i>No. entries</i>	<i>URL</i>
<i>Localisation databases</i>		
OXFORD	5652	<a href="http://www.informatics.jax.org">http://www.informatics.jax.org</a> <sup>a</sup>
MIMMAP	7171	<a href="http://www.ncbi.nlm.nih.gov/omim/">http://www.ncbi.nlm.nih.gov/omim/</a>
MGD	24925	<a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a>
GDB	51917	<a href="http://www.gdb.org/gdb/">http://www.gdb.org/gdb/</a>
<i>Expression and phenotype databases</i>		
SWISSPROT	5908	<a href="http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html">http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html</a>
SWISSNEW	5875	<a href="http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html">http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html</a>
SPTREMBL	23567	<a href="http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html">http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html</a>
REMTREMBL	19036	<a href="http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html">http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html</a>
TREMBLNEW	10394	<a href="http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html">http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html</a>
Medline	58978	<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a> <sup>b</sup>
TBASE	6768	<a href="http://tbase.jax.org/">http://tbase.jax.org/</a> <sup>b</sup>
GXD	24925	<a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a>
OMIM	13099	<a href="http://www3.ncbi.nlm.nih.gov/omim/">http://www3.ncbi.nlm.nih.gov/omim/</a>
MLC	24925	<a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a>
<i>Other database(s)</i>		
GeneCards	20417	<a href="http://bioinformatics.weizmann.ac.il/cards/">http://bioinformatics.weizmann.ac.il/cards/</a>

<sup>a</sup>Accessed after processing from the local mirror site: <http://www.cmbi.kun.nl/srs/>. <sup>b</sup>Accessed directly using this address. Remaining sites were accessed using the local mirror site: <http://www.cmbi.kun.nl/>.

**Table 2** Selected disorders

<i>Syndrome (MIM#)</i>	<i>Expression terms</i>	<i>Genetic localisation</i>
Acro-Dermato-Ungual-Lacrimal-Tooth Syndrome [103285]	Limb/Hand/Teeth	3q27
Alagille Syndrome [118450]	Liver/Eye/Heart	20p12
Hand-Foot-Uterus Syndrome [140000]	Limb/Genital	7p15-7p14.2
Holt-Oram Syndrome [142900]	Limb/Heart	12q24.1
Multiple Synostoses Syndrome 1 [186500]	Ear/Joint	17q22
Noonan Syndrome [163950]	Skeletal/Heart	12q24.1
Renal-Coloboma Syndrome [120330]	Renal/Eye	10q24.3-10q25.1
Townes-Brocks Syndrome [107480]	Limb/Ear	16q12.1
Tricho-Dento-Osseous Syndrome [190320]	Ectoderm/Skeleton/Tooth	17q21.3-17q22
Ulnar-Mammary Syndrome [181450]	Limb/Mammary	12q24.1

found in the syntenic region in mouse, for which the homologous human gene is found to map outside the critical interval, and (4) a list of all the remaining human genes that are present in the genetic interval, but which do not match the expression profile. The data in the HTML files was extracted and converted to a spreadsheet for further analysis.

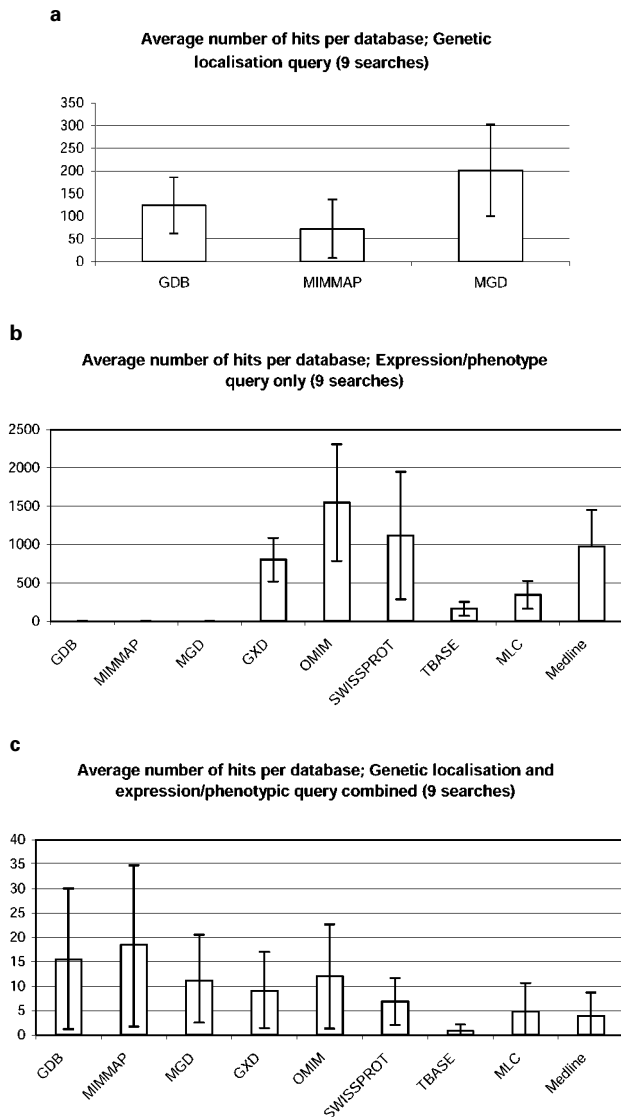
## Results

The evaluation queries were performed in batch in June and July 2001. The processing time of queries using both genetic localisation and expression/phenotypic information varied from 2 min for simple queries to 10 min for complex queries. The number of hits per database for the genetic localisation, expression/phenotypic and the combined queries are presented in Figure 2a,b,c respectively. For the genetic localisation query, no large differences were found between the GDB and MIMMAP. The number of MGD hits is relatively large for several regions. First, because this list

includes both mouse genes whose human homologues are present in GDB as well as a smaller list of mouse genes for which a human homologue could not be identified, either directly or by applying the synonym list. In addition, the conversion through the Oxford grids caused more genes to be retrieved because of wider segment limits. Both Figure 2b and c show a very small contribution from TBASE compared to the other queried databases. This likely reflects the relatively small number of genes in TBASE as well as inconsistent gene naming.

All the causative genes were found in the queries done with only the genetic localisation data (Table 3). The average number of genes in a disease critical interval was 165. This number varied from 322 in the case of Tricho-Dento-Osseous syndrome in 17q21.3-17q22, to only 49 in Townes-Brocks syndrome located at 16q12.1.

Combining genetic localisation with expression/phenotype data was most successful if a Boolean *or* was used to combine expression sites. In all 10 such cases, the causative



**Figure 2** The average number of hits per database in the candidate gene list of the GeneSeeker: (a) only the genetic localisation query; (b) only expression/phenotypic query, combining the search terms with a Boolean *or*; (c) genetic localisation and expression/phenotypic information combined with a Boolean. For the expression terms the thesaurus table is used. The ranges indicate the standard deviation of the mean.

gene was retrieved. Starting from an average number of 165 positional candidate genes (range 49–322), the number of candidate genes that matched both location and expression pattern was reduced to 22 (range 2–63). A match was also obtained for both syndromes for which the gene had not previously been identified as causing the disease. For ADULT syndrome, a candidate gene list of 12 genes was generated, reflecting a 10-fold reduction from 116 positional candidate genes. *TP63*, which was subsequently

been proven to be the causative gene for ADULT syndrome, was present among these 12 selected genes.<sup>11</sup>

A similar result was obtained for Noonan syndrome. Using 'skeletal and heart' as search terms, the number of genes from chromosome band 12q24.1 was 174. This was reduced to 10 in the candidate gene shortlist. Among this final selection was the *PTPN11* gene, which indeed causes Noonan syndrome.<sup>10</sup>

## Discussion

Human disease genes can sometimes be rapidly identified by using information on RNA expression patterns or by studying knockout phenotypes in mice. For instance systematic screens for genes expressed in retina or inner ear are currently being applied successfully in labs around the world in order to identify genes for deafness or blindness respectively.<sup>12–14</sup> Similarly, direct comparison of human and mouse phenotypes allowed for the rapid recognition of *ROR2* as the Robinow syndrome gene.<sup>15,16</sup>

A systematic approach to this conservation of phenotypes has already been attempted and is presented in the Dysmorphic Human-Mouse Homology Database (DHMH).<sup>17</sup> Others have attempted to use cross-species conservation with invertebrates to identify genes that underlie human developmental syndromes and diseases.<sup>18</sup>

All this argues for a systematic bioinformatics approach that includes all available information from public databases to prioritise among positional candidate genes. In a pilot experiment it was previously shown that it is possible to use a bioinformatics approach to identify plausible candidate genes for human multiple congenital anomaly syndromes by systematically using data on murine gene expression patterns.<sup>7</sup>

We have since developed this data mining approach further to create a web-based tool that combines data on genetic localisation from OMIM, GDB, and MGD with data on gene expression from GXD and SWISSPROT/TrEMBL and data on phenotypes in humans (OMIM, Medline) and mice (MLC, TBASE). This approach mimics the steps currently undertaken in most human genetics labs around the world once a critical region for a genetic disease is identified. The biggest advantage of the current automated approach is that it provides combined data from nine databases in a matter of minutes, rather than hours or days if individual databases are queried one gene at a time. Genetic localisation and expression databases were used in almost equal proportion. The number of hits per database varied greatly, but all contributed to the final selection of candidate genes. Of all databases, TBASE contributed the smallest number of genes. This is partly due to the fact that gene names in TBASE often do not conform to the nomenclature used in OMIM or GDB. Moreover, TBASE presently contains information on only a small number of genes.

The most successful search strategy with the GeneSeeker was by using the thesaurus in combination with *or* Boolean

**Table 3** Selectivity and reduction. All queries were performed using expanded thesaurus terms

Syndrome	Localisation selectivity		Expression selectivity		Candidate gene selectivity	
	Total	or	or	and	or	and
Acro-Dermato-Ungual-Lacrimal-Tooth Syndrome	1/116		1/2664	0/109	1/12	0/2
Alagille Syndrome	1/102		1/6435	1/420	1/17	1/3
Hand-Foot-Uterus Syndrome	1/148		1/4550	1/78	1/29	1/2
Holt-Oram Syndrome	1/154		1/4523	1/136	1/18	1/2
Multiple Synostoses Syndrome 1	1/273		1/846	0/102	1/10	0/5
Noonan Syndrome <sup>1</sup>	1/174		1/1090	1/349	1/21	1/12
Renal-Coloboma Syndrome	1/154		1/3774	1/504	1/33	1/7
Townes-Brocks Syndrome	1/49		1/1874	0/74	1/2	0/1
Tricho-Dento-Osseous Syndrome	1/322		1/3046	1/189	1/63	1/4
Ulnar-Mammary Syndrome	1/154		1/4523	0/136	1/18	0/2

The localisation selectivity represents the number of genes in the genetic region, whereas the expression selectivity reflects the number of genes that match the expression terms specified by the user; the candidate selectivity is the combination of the two. The Boolean operators indicate how the primary expression terms (Table 2) are combined. <sup>1</sup>:This analysis was performed in October 2001.

operators. More restrictive strategies failed in a significant proportion of cases, suggesting that the data in the databases is still incomplete, and that inappropriate search terms may have been used that failed to detect the presence of the gene in one or more databases. Some of these failures are to be expected as no single system has been adopted for scoring expression patterns and phenotype across the various databases. This situation is likely to improve considerably over the next few years. First, the genomic databases presently contain only draft versions of the genome with many genes yet to be identified, and properly annotated.<sup>19,20</sup> In addition, efforts are currently underway to set up more complete databases on gene expression and on knockout phenotypes. As one example, a comprehensive inventory of expressed mouse genes during development is in progress.<sup>21</sup> Adding such databases to combined data mining strategies as presented here for the GeneSeeker may further improve their performance. Given the simplicity of the approach that is incorporated in the GeneSeeker tool, one might have expected that other similar applications might already exist. To the best of our knowledge this is not the case. Specifically, no programme appears to be available that evaluates gene expression or phenotype information to aid with selection of positional candidate genes.

It is encouraging that all 10 causative genes were found for the human malformation syndromes with known genesis, and that this was accompanied by an on average 10-fold reduction compared to using localisation data only. We acknowledge that only prospective studies of syndromes that have not yet been defined molecularly can establish the true value of the bioinformatics tool described here. However, a considerable number of human disease genes have already been identified wholly or partly by virtue of comparing their mutant murine phenotypes and expression patterns. This by itself suggests that this approach can only become more effective as more information becomes available for each human gene. The modular setup employed in

this first version of the GeneSeeker should allow easy expansion by adding further databases to improve the detection rate of disease genes. We are currently adding the Unigene database<sup>22</sup> and other EST database sources in order to expand the available information on expression patterns. SAGE (Serial Analysis of Gene Expression) data can also be added in the future, thereby further improving the sensitivity of the tool. Some text-based modules such as Medline may become more effective by using MeSH (Medical Subject Heading) terms and context sensitive searches.

Additional features like a comparison of old and new results, automatic selection of expression terms based on the OMIM clinical synopsis, and the ability to use STS marker data and physical coordinates rather than chromosome bands to specify genetic localisation are currently under development.

In its current form, the GeneSeeker is mainly suited for malformation syndromes in which the assumption can be made that the disease gene has an aberrant or absent gene expression in the affected tissues. For metabolic diseases other strategies can be applied, for example incorporating biochemical pathways such as the Kyoto Encyclopedia of Genes and Genomes (KEGG).<sup>23</sup> This can also be added to future versions of the programme.

In spite of the obvious limitations of the system, even in its present early stage of development the GeneSeeker (version 2.0) offers researchers a useful tool to generate a starting list of candidate genes involved in human genetic disorders. The GeneSeeker site will be continuously updated and curated by the Center for Molecular and Biomolecular Informatics at Nijmegen University. The current setup of the GeneSeeker relies on external databases. This means that regular checks of web addresses and database structures will be necessary to avoid losing individual databases. We intend to provide such regular follow-up, and note that the setting within a centre that provides support for more than 70 databases already would seem to be ideal for this. (Average downtime for these databases over the past 3 years

has been less than 0.5%.) Also, the current GeneSeeker system has a number of advantages. Using external databases means that we avoid data warehousing. Therefore, all data are up to date and we would argue that in practice WWW front-ends are more stable than their underlying relational tables. In fact when changing the internals of the database, database developers often try to keep the WWW front-end unchanged. In conclusion, current developments in the availability of genomics data as well as improving bioinformatics strategies support the notion that data mining approaches as applied in the GeneSeeker may become a useful adjunct to wet lab experiments in human genetics.

### Note added in proof

The programme described here has previously been presented at scientific conferences and in abstracts as the 'GeneMachine'.<sup>24</sup> In order to avoid confusion with a recently published programme,<sup>25</sup> we henceforth shall use the name GeneSeeker.

### Acknowledgements

Partly supported by grants from N.W.O./Unilever (grant number 326756 to JAM Leunissen) and from the Irene kinderziekenhuis Foundation (to HG Brunner).

### References

- 1 Letovsky SI, Cottingham RW, Porter CJ, Li PW: GDB: the Human Genome Database. *Nucleic Acids Res* 1998; **26**: 94–99.
- 2 Blake JA, Eppig JT, Richardson JE, Bult CJ, Kadin JA: The Mouse Genome Database (MGD): integration nexus for the laboratory mouse. *Nucleic Acids Res* 2001; **29**: 91–94.
- 3 Bairoch A, Apweiler R: The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 2000; **28**: 45–48.
- 4 Ringwald M, Eppig JT, Begley DA *et al*: The Mouse Gene Expression Database (GXD). *Nucleic Acids Res* 2001; **29**: 98–101.
- 5 Woychik RP, Wassom JS, Kingsbury D, Jacobson DA: TBASE: a computerized database for transgenic animals and targeted mutations. *Nature* 1993; **363**: 375–376.
- 6 Stern AM, Gall Jr JC, Perry BL, Stimson CW, Weitkamp LR, Poznanski AK: The hand-foot-uterus syndrome: a new hereditary disorder characterized by hand and foot dysplasia, dermatoglyphic abnormalities, and partial duplication of the female genital tract. *J Pediatr* 1970; **77**: 109–116.
- 7 van Steensel MA, Celli J, van Bokhoven JH, Brunner HG: Probing the gene expression database for candidate genes. *Eur J Hum Genet* 1999; **7**: 910–919.
- 8 Propping P, Zerres K: ADULT-syndrome: an autosomal-dominant disorder with pigment anomalies, ectrodactyly, nail dysplasia, and hypodontia. *Am J Med Genet* 1993; **45**: 642–648.
- 9 Jamieson CR, van der Burgt I, Brady AF *et al*: Mapping a gene for Noonan syndrome to the long arm of chromosome 12. *Nat Genet* 1994; **8**: 357–360.
- 10 Tartaglia M, Mehler EL, Goldberg R *et al*: Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nat Genet* 2001; **29**: 465–468.
- 11 Amiel J, Bougeard G, Francannet C *et al*: TP63 gene mutation in ADULT syndrome. *Eur J Hum Genet* 2001; **9**: 642–645.
- 12 Dryja TP: Gene-based approach to human gene-phenotype correlations. *Proc Natl Acad Sci USA* 1997; **94**: 12117–12121.
- 13 den Hollander AI, van Driel MA, de Kok YJ *et al*: Isolation and mapping of novel candidate genes for retinal disorders using suppression subtractive hybridization. *Genomics* 1999; **58**: 240–249.
- 14 Blackshaw S, Fraioli RE, Furukawa T, Cepko CL: Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. *Cell* 2001; **107**: 579–589.
- 15 van Bokhoven H, Celli J, Kayserili H *et al*: Mutation of the gene encoding the ROR2 tyrosine kinase causes autosomal recessive Robinow syndrome. *Nat Genet* 2000; **25**: 423–426.
- 16 Afzal AR, Rajab A, Fenske CD *et al*: Recessive Robinow syndrome, allelic to dominant brachydactyly type B, is caused by mutation of ROR2. *Nat Genet* 2000; **25**: 419–422.
- 17 van Steensel MA, Winter RM: Internet databases for clinical geneticists—an overview. *Clin Genet* 1998; **53**: 323–330.
- 18 Banfi S, Borsani G, Rossi E *et al*: Identification and mapping of human cDNAs homologous to Drosophila mutant genes through EST database searching. *Nat Genet* 1996; **13**: 167–174.
- 19 Venter JC, Adams MD, Myers EW *et al*: The sequence of the human genome. *Science* 2001; **291**: 1304–1351.
- 20 Lander ES, Linton LM, Birren B *et al*: Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.
- 21 Davidson D, Bard J, Kaufman M, Baldock RA: The Mouse Atlas Database: a community resource for mouse development. *Trends in Genetics* 2001; **17**: 49–51.
- 22 Schuler GD, Boguski MS, Stewart EA *et al*: A gene map of the human genome. *Science* 1996; **274**: 540–546.
- 23 Kanehisa M, Goto S, Kawashima S, Nakaya A: The KEGG databases at GenomeNet. *Nucleic Acids Res* 2002; **30**: 42–46.
- 24 Brunner H, Cuelenaere K, Kemmeren P, van Driel MA, Leunissen JAM: The Genemachine: A tool for the extraction and integration of information from web-based genetic databases. *Eur J Hum Genet* 2000; **8**: 130.
- 25 Makalowska I, Ryan JF, Baxeavanis AD: GeneMachine: gene prediction and sequence annotation. *Bioinformatics* 2001; **17**: 843–844.