Part 1: Previous Track Record

Personnel

Dr. Casey Bergman is a recently-appointed lecturer in the Faculty of Life Sciences at the University of Manchester, with a primary affiliation in Bioinformatics and Functional Genomics. The major focus of his research group is to understand the function, organisation and evolution of genome sequences using computational methods. Dr. Bergman's research achievements include development of the prevailing model for *cis*-regulatory evolution of binding site turnover under stabilising selection [1], performing the pilot analysis to aid the choice of additional *Drosophila* genomes to be sequenced [2], and providing the official FlyBase annotation of transposable elements and transcription factor binding sites in *Drosophila* [3, 4]. Since 2000, Bergman has published 23 peer-reviewed articles in journals such as *Nature*, *PLoS Biology*, *PNAS, Genome Research*, *Genome Biology* and *Bioinformatics*, and has been invited to give over 30 research seminars in the US, UK and Europe. Dr. Bergman is also a member of the Royal Society International Policy Committee, a contributing member of the Bioinformatics subsection of the Faculty of 1000, co-leader of the NSF-funded National Center of Evolutionary Synthesis Working Group on *cis*-regulatory evolution, and was the only UK project leader in the *Drosophila* 12 Genomes Consortium [5].

Dr. Goran Nenadic is a Lecturer in Text Mining in the School of Computer Science, University of Manchester. He is also an associated academic member of the National Centre for Text Mining (NaCTeM) and the Manchester Interdisciplinary BioCentre (MIB). His research interests are focused on automatic information and knowledge extraction from biological literature. Dr Nenadic leads the BBSRC-funded project "Mining term associations from literature to support knowledge discovery in biology," [1] whose aim is to develop sophisticated text mining and machine learning techniques to assist biologists in hypotheses generation. Previously, Dr. Nenadic was involved in the development of text-mining methods for recognition and warehousing of biomedical terminology (EUREKA Bio-PATH project) and prediction of protein function (BBSRC-funded "Protein function classification using text data mining"). His group was one of two UK groups that took part in the first evaluation of protein functional annotation from text (BioCreative I, task 2). With the EBI and NaCTeM, Dr. Nenadic coordinates efforts for interoperability in text mining in the biological domain [6]. He was part of the team of that was awarded a Daiwa Adrian Prize for "Knowledge Mining from Biology Texts" in 2004 (a triennial award for exceptional collaboration between UK and Japanese research teams in sciences).

Recent Work

Dr. Bergman's recent work in integrative genome bioinformatics and comparative genomics has included major contributions to the annotation and analysis of *cis*-regulatory and transposable elements in the *Drosophila* genome [3-5]. His interest in integrating genome bioinformatics and text-mining arose from a long-term effort to manually curate *Drosophila* transcriptional regulatory data from primary text, which culminated in the release of the *Drosophila* DNAse I footprint database [4], the first open-access database of transcription factor binding sites mapped to genome coordinates. The importance of mapping these critical sequence data from the primary literature to genome coordinates is indicated by the fact that this work was quickly incorporated into the official FlyBase genome annotation, and imported into the Ensembl, UCSC, FlyMine, Transfac and Open Regulatory Annotation (ORegAnno) databases within a year of publication. Based on this experience with the labor-intensive manual curation and warehousing of *cis*-regulatory data [4, 7], Dr. Bergman has been recently involved in efforts to develop automated text-mining technologies to support research in transcriptional regulatory biology (see **Case for Support Part 2**), including co-organising the first RegCreative Jamboree. This focused workshop held in Nov 2006 brought together 45 biologists, bioinformaticians and text miners from the UK, Europe and North America to outline requirements and aspects of the regulatory annotation process that could be automated using text-mining approaches. The RegCreative Jamboree demonstrated the need for text-mining systems in genome informatics. Dr. Bergman was asked to present the outcomes of the RegCreative Jamboree at the Second BioCreative Challenge Text-Mining Workshop in April 2007, as a means to motivate the regulatory bioinformatics challenge to the wider text-mining community.

Current research in Dr. Nenadic's group is focused on automated but customized extraction of knowledge from text to support specific biological curation and research tasks. The generic approach is integrates extensive terminological processing of the literature and machine learning for predicting roles and functions of biological entities based on their textual features [8]. The group has developed a novel approach for harvesting contextual descriptions to characterize biological entities and classes of interest [9]. These descriptions are used as features in machine learning/data mining to predict functional classes and roles (e.g. automatic prediction of human protein functional annotations using GO-terms [10]; identification of transcription factors [11]). Dr. Nenadic also investigates building controlled vocabularies using automated recognition (now provided as a Web service through the National Centre for Text Mining), classification and identification of terminology [12, 13]. Dr. Nenadic is also co-organiser of the international workshop on "Building and evaluating resources for biomedical text mining" to be held in May 2008.

Research Environment

The RAE 5*-rated Faculty of Life Sciences (FLS) at the University of Manchester is ideally suited for bioinformatics-led genomics research, and is home to the largest academic grouping in bioinformatics at any UKHEI. FLS research in

[1] An interim report for this grant is attached to this application.

Bioinformatics and Functional Genomics spans the full range of from genome to systems biology, and provides the hub to the over 100 active researchers in bioinformatics spread across the University in Life, Physical and Computer Sciences. A 32-node, dual-core Beowulf cluster is currently dedicated to bioinformatics research in FLS and is supported by a full-time system administrator; a second 32-node high performance computing cluster is being built using funds from a recent ~£170,000 BBSRC REI award. The Bergman lab is located in the newly built Michael Smith Building and currently consists of 1 PDRA (Dr. Ian Donaldson, analyzing ChIP-Chip tiling array data, jointly supervised with Prof. Andy Sharrocks, University of Manchester) and 3 postgraduate students (Michael Barton – yeast systems biology, Paul Cartwright – human genome evolution, and Raquel Linheiro – *Drosophila* transposable elements). The research and training environment in FLS is highly interactive and is facilitated by research seminar series in Bioinformatics and Functional Genomics with internal and external speakers from across the UK and Europe.

The School of Computer Science is the first Computer Science department in the UK and has remained an internationally recognised centre of excellence ever since, and was awarded a 5* in the 2001 RAE as one of the top six departments in the UK. A key theme across the School's activities is the interdisciplinary nature of our research, focusing on providing frameworks and advanced technologies (e.g. data analysis and integration, e-science, Semantic Web) for other disciplines, in particular for medicine, biology and chemistry. The School's Text Mining and Natural Language Processing research group has developed a large variety of expertise in extracting knowledge from textual data, in particular in the biomedical domain. Particular strengths include collection and standardisation of language resources (lexicons, terminologies, ontologies), automatic terminology extraction, ontology-driven information extraction, knowledge representation techniques, and machine learning. The Nenadic lab consists of four members: one PDRA (Dr. Hui Yang, working on machine-learning based knowledge extraction) and three postgraduate students (Mark Greenwood – topic-focused information retrieval, Hammad Afzal – term classification, and Farzaneh Sarafraz – integrative text/data mining). The group is ideally located in the Manchester Interdisciplinary Biocentre (MIB), which aims to apply novel technologies derived from the earth and physical sciences to biological problems and provides state-of-the-art facilities for bioinformatics and systems biology research. The School and the group host and lead the UK National Centre for Text Mining (NaCTeM, co-located in the MIB), which is the first publicly funded text-mining centre in the world.

Strong links exist between the Faculty of Life Sciences and School of Computer Science have been built through joint appointments, collaborative projects, post-graduate programs, research co-supervision, seminars, and centres. The strengths of the individual groups and supporting research infrastructure mean that there can be no better academic environment in which to execute the work proposed in this application.

References

1. Ludwig, M.Z., C. Bergman, N.H. Patel, and M. Kreitman (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**: 564-7.
2. Bergman, C.M., B.D. Pfeiffer, D.E. Rincon-Limas, R.A. Hoskins, et al. (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol* **3**: RESEARCH0086.
3. Quesneville, H., C.M. Bergman, O. Andrieu, D. Autard, et al. (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol* **1**: e22.
4. Bergman, C.M., J.W. Carlson, and S.E. Celniker (2005) *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics* **21**: 1747-9.
5. Clark, A.G., M.B. Eisen, D.R. Smith, C.M. Bergman, et al. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203-218.
6. Rebholz-Schuhmann, D., H. Kirsch, and G. Nenadic. (2006) *IeXML: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules*. in *Proceedings of Joint BioLINK and Bio-Ontologies SIG Meeting, ISMB 2006*.
7. Montgomery, S.B., O.L. Griffith, M.C. Sleumer, C.M. Bergman, et al. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics* **22**: 637-40.
8. Nenadic, G., I. Spasic, and S. Ananiadou (2003) Terminology-driven mining of biomedical literature. *Bioinformatics* **19**: 938-943.
9. Nenadic, G. and S. Ananiadou (2006) Mining semantically related terms from biomedical literature. *ACM Transactions on ALIP* **5**: 1-22.
10. Rice, S., G. Nenadic, and B. Stapley (2005) Mining protein function from text using term-based support vector machines. *BMC Bioinformatics* **6(Suppl 1)**: S22.
11. Yang, H., G. Nenadic, and J. Keane (2008) Identification of transcription factors contexts in literature using machine learning approaches. *BMC Bioinformatics* **(in press)**.
12. Rebholz-Schuhmann, D., H. Kirsch, S. Gaudan, M. Arregui, et al. (2006) *Annotation and disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition*. in *Proceedings of NLPXML 2006, EACL 2006*.
13. Yang, H., G. Nenadic, and J. Keane (2007) A cascaded approach to normalising gene mentions in biomedical literature. *Bioinformation* **2**: 197-206.

Part 2: Description of the Proposed Research and its Context

Background

Advances in DNA sequencing technology over the last 20 years have drastically increased the rate of production of genomic sequence data, which in turn has directly accelerated the rate of biological discovery and publication. Genome sequences and their supporting computational annotations are well-served by generic genome portals, such as the Ensembl or UCSC browsers, or by specific model organism databases, such as FlyBase. More recently, effort has been invested in integrating genomic database resources with other gene-based microarray or protein resources, such as ArrayExpress or UniProt, which are now highly interoperable with genome or model organism databases. In contrast, essentially no effort has been made to systematically integrate genome sequences directly with the biological literature, despite the fact that these are the two most heavily relied-upon sources of information for many biologists. The ability to navigate directly between genomes and the biomedical literature, and to perform cross- and multi-lingual queries using both textual and genomic constraints would greatly aid experimental and computational researchers alike, and would provide a unique and much-needed bridge between two of the fastest growing sources of biological information. Here we propose to develop an open-access resource called pubmed2ensembl that links biological literature directly to genome sequences, allowing integrated queries over genomic and textual information *via* human and programmatic web interfaces.

The primary portal for access to biological literature by most biologists is *via* searching abstracts warehoused at the US National Library of Medicine's PubMed database. Abstracts of new articles are submitted to PubMed by journals, given a PubMed identifier (PMID), and then tagged by human curators with Medical Subject Heading (MeSH) terms that consist of a controlled vocabulary of scientific concepts. Although names for some important genes are included as part of the MeSH controlled vocabulary, the assignment of specific gene or genomic regions to PMIDs is not currently a part of the standard warehousing of the biological literature. It would be trivial process to map papers to genes or genomes if authors only used standardised genome database identifiers to refer to gene names in text. Unfortunately, this is not the case, and genes are often referred to in the literature by many variant spellings or synonyms. Major advances have been made over the last few years in the field of automated gene name recognition and mapping of gene names to database identifiers [1], however these text-mining systems have not yet been integrated into the Medline curation process. As a result, thousands of primary research articles in PubMed that refer to specific regions in sequenced organisms are not directly linked to any genomic data.

For a limited number of model organisms (such as yeast, *Drosophila*, *C. elegans*, *Arabidopsis* and mouse) dedicated teams of human curators scan the biological literature for publications that contain references to species-specific gene names, and subsequently add PMIDs to gene records in model organism databases (MOD) such as SGD, FlyBase, WormBase, TAIR and MGI. However, the ever-increasing rate of publication often necessarily creates a backlog of uncurated articles, or requires MOD curation teams to adopt triage procedures whereby only a subset of the literature is curated and linked to gene and genome resources. More importantly, thousands of articles may refer to genes in sequenced genomes for species with no such dedicated literature curation staff, and may never be mapped to genes or genomic regions. Even for those genomes supported by dedicated curation efforts, PMIDs are typically linked to internal identifiers (e.g. FlyBase uses internal FBrf identifiers), which are linked to gene identifiers, which are in turn linked to genome coordinates. Thus in the case of FlyBase, the user must execute 3-4 links (if these are available) to go from a gene- or genomic region to a PubMed record. Likewise, at least two links are needed to go from Ensembl genome coordinates to genes to PMIDs found in external database records (e.g. UniProt). The same is true for the US National Center for Biotechnology Information (NCBI) Entrez Gene project, which integrates data from external curated resources (e.g. FlyBase) with data from GenBank records associated with publications to provide links from RefSeq genes to PMIDs. Importantly, while these curated data provide a rich set of gene-PMID links that can be exploited by the proposed system, no resource currently provides the capability of performing text-based queries on curated document sets restricted by gene or genomic regions. Providing an integrated resource between genomes and the literature will have a major impact on biological research, reciprocally allowing improved access to the literature for specific genomic intervals as well as improved analysis of genomes based on functional data in the biomedical literature.

Aims and Objectives

We aim to overcome the lack of integration between genome sequences and biological literature by mapping all possible PMIDs to regions of all sequenced genomes supported by the Ensembl database, using the combination of approaches diagrammed in **Figure 1**. The pubmed2ensembl resource will be constructed in a 12-month period by completing four primary objectives, each to be executed as a separate Work Package (WP):

- WP1: Construction of the pubmed2ensembl relational database and population with external data.
- WP2: Mapping of PubMed corpus to Ensembl genomes using text-based sequence extraction (text2seq).
- WP3: Deployment of the pubmed2ensembl server and integration with external databases.
- WP4: Development of pubmed2ensembl human and programmatic web interfaces.

We will use several sources of both human-curated and automatically-extracted gene-PMID links to populate the pubmed2ensembl database, attach evidence codes to each mapping to allow user-based quality filters, and provide both human and programmatic web interfaces to access the data. Queries to the pubmed2ensembl system can be executed using genome- or text-based data types and return data types in the same or complementary domain. The capability for such cross- and multi-lingual queries over text and genomic data will be a novel and defining feature of the pubmed2ensembl system (for examples, see **Use Cases** below). We will heavily rely upon established bioinformatics resources (such as the Ensembl Genome Browser [2], the Distributed Annotation System [3], BioMart [4] and NCBI eutils [5]) to build and connect components of pubmed2ensembl in order to speed-up development time, allow maximal interoperability with other

systems, and to allow users to capitalise on their familiarity with these widely-used systems. Using Ensembl genomes focuses our system on a subset of highly-supported, well-studied genomes and critically will allow cross- or multi-species analysis of all gene-PMID mappings using the Ensembl comparative genomics database [2]. pubmed2ensembl will uniquely capitalise on the novel demonstration by the PIs that DNA sequences can be extracted from full-text articles and mapped to genomes with high precision (see **Work by the Applicants Leading up to the Current Project** below). Finally, the pubmed2ensembl resource will be open-access and open-source so as to maximise access to it by the wider research community, and to encourage use of pubmed2ensembl data or services in other text mining or genome informatics projects.

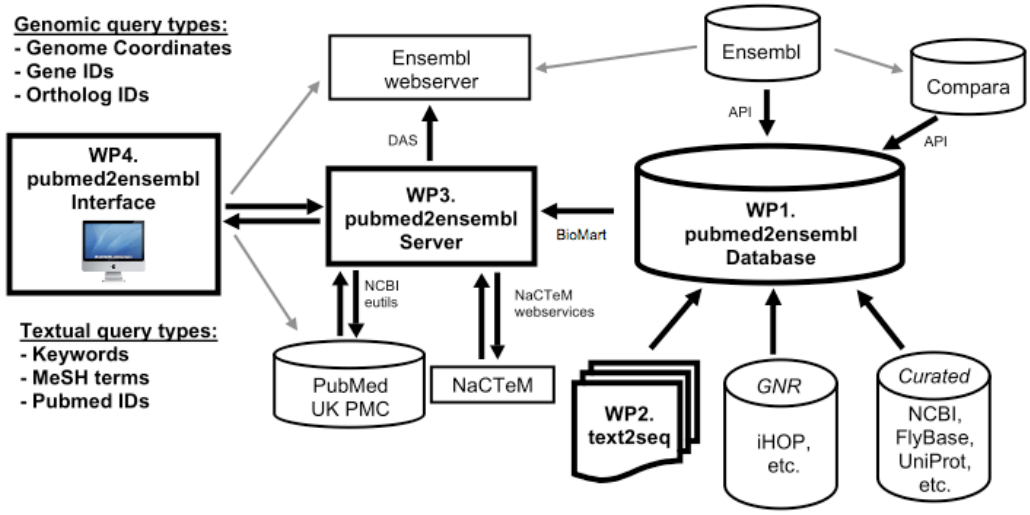

Figure 1. Overview of pubmed2ensembl system architecture. Components of pubmed2ensembl that will be developed in Work Packages 1-4 of this project are highlighted in bold. Rectangles indicate servers or workstations, cylinders indicate databases, arrows represent software interfaces or utilities, and multi-copy documents represent full-text articles.

Use Cases

We have identified four general types of use cases, many of which will be completely novel to the pubmed2ensembl system. To illustrate them, we use the example of Alice, a biologist who wishes to study genes involved in sexual dimorphism in mice. Sexual dimorphism, like other important biological processes, is not included in the Gene Ontology (GO) and thus cannot be used to query genomic data directly. Alice would like support for the following integrated querying services:

1) Text querying constrained by genomic data. Alice wishes to find papers that report genes involved in sexual dimorphism on the X-chromosome. She executes a "sexual dimorphism" query using the pubmed2ensembl system, restricting her search to papers on the X-chromosome interval, and in the homologous intervals in all mammalian genomes.
2) Genomic querying constrained by text. Alice has conducted a microarray experiment to find genes involved in sexual dimorphism. To calibrate her analysis and guide further experiments, she loads her gene list into the pubmed2ensembl system to identify genes that have (and do not have) published evidence of being involved in sexual dimorphism.
3) Text querying resulting in genome data. Alice wishes to obtain the GO terms that are frequently linked to genes involved in sexual dimorphism in mouse. She executes the text query "sexual dimorphism" at pubmed2ensembl, the system retrieves all genes from papers that match the query term and provides automated retrieval of GO terms associated with these genes.
4) Genomic querying resulting in text data. Alice wishes to identify candidate functions for a list of genes and their homologs to complement GO annotations. She uploads the list of genes to the pubmed2ensembl system, which retrieves all papers for these genes and their mammalian orthologues, and then submits them to NaCTeM's Termine webservice for automated recognition of over-represented terms.

Related Work and Justification for the Current Resource

Very few systems have attempted to integrate genomic and textual data, none of which fully achieve the aims that we propose in this project. MedBlast [6] is a web-based tool that finds PubMed articles associated with a query sequence by direct look-ups for PMIDs and extraction of gene names from CDS fields in GenBank records obtained by the BLAST search. MedBlast aims to find publications related to an input sequence, but does not provide an integrated resource of genome and biological literature, does not allow multi-lingual or genome-based (i.e. gene name, coordinate) queries, and does not allows the user to navigate from PMIDs to genome data. Entrez Gene [7] provides ~4 million high quality links from genes to PMIDs based largely on external sources of curated data (e.g. FlyBase) and GenBank records, but critically does not allow gene- or genome-based constraints on text queries to the PubMed database, nor does it utilise the extremely rich set of gene-PMID links that can be generated using gene name recognition or text-based sequence extraction. The iHOP system [8] uses gene name prediction in PubMed abstracts to generate hyperlinks based on gene names to navigate the biomedical literature. It provides the largest open-access dataset of predicted gene-PMID links, but does not allow genome-based constraints to access information or

execute PubMed queries. CiteXplore (unpublished) provides enhanced human and programmatic text-based queries to PubMed records, and cross-references to UniProt, InterPro, SwissProt/Trembl records in the resulting documents set. CiteXplore does not link directly to Ensembl genes, and therefore does not provide cross- or multi-lingual queries or navigation capabilities. The PosMed system (unpublished) is the closest system to the one that we propose in that it allows PubMed queries that are restricted to specific genes or intervals, but only in 3 genomes (human, mouse or rat) and does not utilise comparative genomic information or provide a programmatic interface. PosMed links genes to PMIDs using an unspecified gene name recognition system based on term frequency, but does not use curated gene-PMID links or gene-PMID links based on sequences extracted from text. PosMed is also limited by its tight integration with the Riken 'Omics browser, reliance only on probabilistic gene name recognition, and does not allow the user to navigate from genomes to PMIDs.

Although MedBlast, Entrez Gene, iHOP, CiteXplore and PosMed demonstrate the emerging need and utility for resources like `pubmed2ensembl`, currently none of these systems provide all of the key features we propose, including: 1) the ability to navigate directly from genomes to PMIDs *and* PMIDs to genomes, 2) the use of both curated *and* predicted gene-PMID links, 3) the integration of comparative genomic searches across the biomedical literature, 4) the availability of human *and* programmatic web interfaces, and 5) the integration with open-source/open-access web technologies. By overcoming the weaknesses of these related systems, our `pubmed2ensembl` system will provide the first fully integrated genome- and text-based query system. The need for such a system is illustrated by the fact that executing even simple cross- or multi-lingual queries such as those outlined in the **Use Cases** (above) are not currently possible without dedicated bioinformatics support.

Preliminary Work by the Applicants Leading up to the Current Project

A major outcome of the interaction between text-mining and genome bioinformatics communities at the RegCreative Jamboree (see **Case for Support Part 1** for details) is a manuscript currently in revision at *Genome Biology* entitled "Text-mining assisted regulatory annotation," with Dr. Bergman as senior author. In this study, Dr. Bergman and colleagues applied an information retrieval system based on a vector space model of MedLine abstracts [9] to identify papers with high *cis*-regulatory content based on a training set of papers with curated *cis*-regulatory data in the ORegAnno Database. Using this model to identify a corpus of PMIDs with relevant data, we demonstrated that DNA sequences can be extracted from full-text articles and mapped to eukaryotic genomes as a novel means to semi-automatically annotate *cis*-regulatory sequences. **Figure 2** provides an example of *cis*-regulatory sequences from 3 full-text articles (PMIDs) that have been automatically extracted from full text articles and accurately mapped to the promoter region of the *hsp70* gene in *Drosophila*. While previous work has shown that DNA sequences can be easily discriminated from English words in biomedical text [10], this work is the first demonstration that DNA sequences can be automatically extracted from biological literature and directly used to link PMIDs to specific genomic regions. This proof-of-principle in the domain of regulatory annotation represents a new hybrid approach to text-based genome annotation and will be generalised in this project to all PMIDs for which we can obtain sequences from full-text articles.

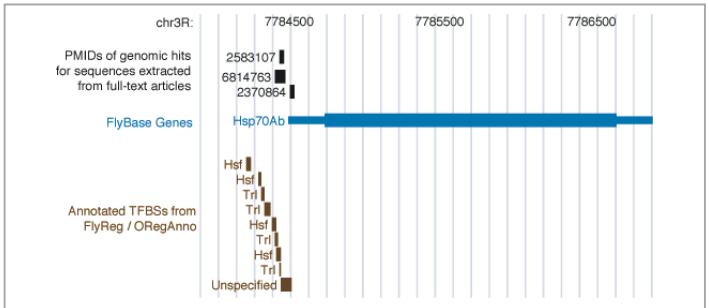

Figure 2. Example of gene-PMID mappings based on DNA sequences extracted from full-text articles. Note that extracted sequences (top, black) map precisely to annotated *cis*-regulatory elements (bottom, brown) in the *Drosophila Hsp70Ab* gene (middle, blue).

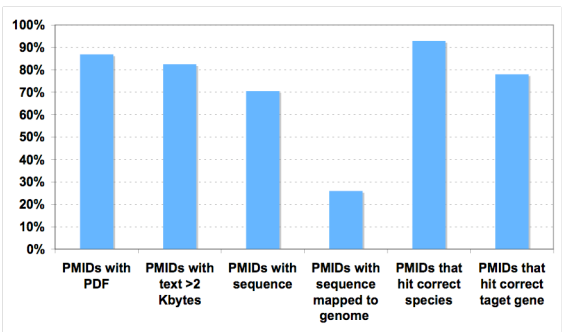

Figure 3. Performance of mapping PMIDs to genes and species using sequences extracted from full text articles. Data are for an initial corpus of 11,437 PMIDs mapped to five genomes (*Drosophila*, *C. elegans*, human, mouse and rat).

The performance of our system in terms of document retrieval, text conversion, sequence extraction, genome mapping, species identification and gene identification on an initial corpus of 11,437 PMIDs is shown in **Figure 3**. Numbers represent averages across mappings to only five model organism genomes (*Drosophila*, *C. elegans*, human, mouse and rat) and therefore represent an underestimate of the performance if applied to the full set of available genome sequences. Our methods to automatically harvest PDFs using a combination of NCBI eutils and the `PERL mechanize` module allow PDFs to be obtained for >85% of PMIDs using the University of Manchester site licences. We find that >82% of PMIDs provide full-text files with non-trivial content (>2 Kbytes) after text conversion. Over 70% of PMIDs in our corpus have DNA-like strings, although many of these are too short to be mapped accurately and uniquely to genome coordinates. Overall, ~25% of PMIDs contains sequences that can be mapped to these five genomes. Nevertheless, on the full PubMed corpus of ~17 million PMIDs, this will potentially translate to thousands of PMIDs that can be mapped to genomes using this approach. Assignment of genomic hits to the nearest gene results in correct gene name identification at rates (>75%) that are comparable to the best gene name recognition software [1]. These genome mappings provide high quality gene-PMID links and represent a unique source of data that will not be available elsewhere. The in-house pipeline we developed for this purpose is robust and scalable and will form the basis of the `text2seq` module that will be released as part of this project (see **WP2** below).

Programme and Methodology

pubmed2ensembl will be developed in four Work Packages, each of which will utilise established open-access bioinformatics resources (such as the Ensembl Genome Browser, the Distributed Annotation System, BioMart, iHOP webservices, NCBI eutils and NaCTeM webservices) to build or connect components of pubmed2ensembl. Re-use of existing resources will speed-up development time, allow maximal interoperability with other systems, and build on the user's familiarity with these systems. Given the experience and track record of the PIs in genome informatics and text mining, together with these existing open-access computational technologies/resources, each objective/WP is rapidly achievable, and will lead to an internationally-recognised, high-impact resource that provides novel integration between genomic data and the biomedical literature. WPs will be conducted in parallel where possible to accommodate long computation times (e.g. for extracting and mapping sequences in WP2) or to facilitate naturally synergistic activities (e.g. server and interface development in WP3 and WP4).

Work Package 1: Construction of the pubmed2ensembl database and population with existing data.

In months 1-3 of the project, we will design a schema and develop a mySQL database (deliverable D1) that will store the basic relational entities that underlie the pubmed2ensembl system: gene-PMID pairs, plus their source codes, and confidence scores. Curated gene-PMID links will be given the higher confidence scores than predicted gene-PMID links based on gene name recognition or text based sequence extraction, which will be given equivalent confidence scores since their performance is comparable. Gene-PMID links with multiple sources of evidence will be given higher confidence scores. Source codes and confidence scores will all be provided to the users as filters in either genome or text-based queries.

In the first instance, we will populate key tables with existing curated and predicted gene-PMID data from external sources (deliverable D2) to allow rapid development and release of the rest of the pubmed2ensembl system. Relevant data will be obtained directly from Ensembl, NCBI, MODs and secondary sequence databases; conversion tables internal to the pubmed2ensembl system will be generated using custom utilities. Data will be imported first from external databases that have used human curation to produce highly reliable gene-PMID pairs. These external databases will include model organism databases (e.g FlyBase, MGI) secondary sequence databases (e.g UniProt) and Entrez Gene. Based on Entrez Gene data alone, we estimate ~1,000,000 curated gene-PMID pairs can be imported from external sources to one of the 37 Ensembl genomes. In contrast to Entrez Gene, we will preserve the original data source so that users can trace the provenance of the curated data. Next, we will import all gene-PMID pairs from the iHOP database, the largest open-access source of automated gene name recognition available for PubMed abstracts. Data import will be facilitated and maintained up-to-date by capitalising on the newly implemented iHOP webservices [11]. Gene-PMID links from iHOP have not been subject to human curation, however the estimated precision of gene name recognition for this system is 94% [12], indicating that these mapping are also of very high quality. iHOP also provides qualitative assessment of the "quality" of each gene name prediction, which we will inherit and maintain to allow users to filter iHOP based gene-PMID links on quality scores.

Our database schema will be designed to extend Ensembl core and comparative (compara) databases. In the first instance, we will import the latest version of the entire core and compara databases for each the 37 species supported by Ensembl. Wrapping these tables together with gene-PMID tables specific to pubmed2ensembl will allow the rapid deployment of a fully functional BioMart (see **WP3** below) that integrates genomic and literature data in an explicitly comparative context. Integration of comparative genomic data will give pubmed2ensembl users the extremely powerful ability to execute textual queries not just for a specific gene or genomic region, but also for all of its homologous sequences in the compara database. This unique design principle is not found in any related system (see above), and closely models the natural workflow in the life sciences, since nearly all biologists leverage the power of comparative data from model organisms in their scholarship and research. The completion of WP1 (Month 3) will be rapid and low-risk but will already lead to a high-impact deliverable from the project. Nevertheless, careful design principles will be followed to ensure reliability and long-term maintenance of the database and external data import utilities, which will form the essential framework for the rest of the project.

Work Package 2: Mapping of the PubMed corpus to genome sequences using text2seq.

Concurrent with **WP1** and extending until complete evaluation of all PMIDs with available PDFs, we will launch our text2seq pipeline to harvest all available full-text articles and perform text-based sequence extraction and genome mapping. We expect that this straightforward but computationally intensive process will lead to hundreds of thousands of high-quality gene-PMID links that will not be available from any other database. We will perform this process in reverse chronological order of publication date to enrich for papers with available full-text articles, and also implement routines to harvest newly published PDFs on a monthly basis. We will execute BLAST searches on all available Ensembl genomes and store the extracted sequences, BLAST score, location and closest gene for each genome hit.

Our preliminary results (see **Work by the Applicants Leading up to the Current Project** above) indicate that ~25% of PMIDs will have sequences that can be extracted and provide gene-PMID links, which is comparable to the proportion of abstracts with identifiable names using automated gene name recognition [13]. We expect that target gene name identification by text2seq will exceed the current levels of ~75% precision, since our preliminary results are based largely on noncoding regulatory sequences, while many of the sequences in full-text will correspond directly to cDNAs or fragments of gene sequences. Preliminary analysis were performed using BLAST to execute genome mappings, but we will investigate alternative fast similarity searches such as ssaha [14] and BLAT [15] to improve the speed and/or accuracy of text2seq. We will also optimise our gene assignment procedure (currently we assign PMIDs to the closest gene to the BLAST hit), which can make incorrect assignments because of nested/overlapping/closely spaced genes.

Based on our successful application of sequence extraction in the regulatory annotation domain for full-text articles that span the 1980s to the present, we expect that ~80% of PMIDs will have PDFs that we can automatically download using NCBI eutils with current University of Manchester site licences. Since we are only linking PMIDs to genes (as is done by the curation teams at FlyBase of NCBI) and not re-distributing text or PDFs, our approach will not violate intellectual property or copyrights. We will investigate harvesting documents via UK PubMed Central (UKPMC) as a complementary strategy that may increase access or coverage of full-text articles. We are also exploring mechanisms to establish link-outs from UKPMC to pubmed2ensembl or to Ensembl *via* pubmed2ensembl (see **letter of support from Vic Lyte**, MIMAS, UKPMC Development Manager), which would substantially increase the visibility and use of pubmed2ensembl and add a unique added value to UKPMC.

Finally, we believe the text2seq application will be of general use to many researchers and we will re-write our current pipeline into a distributable open-source package for users to download and run locally. We will also develop a GUI for users to upload PDF articles individually or in batch for text conversion, sequence extraction and genome mapping. The text2seq package will form a core deliverable of WP2, with a first version to be released by Month 6 (deliverable D3). Only a small amount of full-time human effort is required in this WP2, since the existing text2seq pipeline is fully-functional on a 32-node, dual-core Beowulf cluster, and can easily be integrated with the pubmed2ensembl database. Therefore, aside from packaging the text2seq method for distribution, only limited human intervention will be needed to manage automated updates to the database as papers are processed by the pipeline and submitted to the database (deliverable D4).

Work Package 3: Deployment of the pubmed2ensembl server and integration with external databases and services.

We will establish a dedicated server to run the pubmed2ensembl database and query interfaces. In the first instance, we will deploy pubmed2ensembl by building a BioMart data warehouse [4] around the pubmed2ensembl relational database. BioMart provides a simple and powerful open source platform to create and maintain an advanced query interfaces to relational databases, and it is specifically designed to facilitate data mining of complex datasets such as biomedical text and genomic data. BioMart was designed around the Ensembl schema and is therefore naturally suited to our database design, which extends the Ensembl schema. BioMart provides several alternative interfaces to a database packaged as a Mart – *via* the command line through martshell, *via* a web browser through martview, and *via* web services using martservice. Thus, by instantiating pubmed2ensembl through BioMart we will concurrently deploy several well-established interfaces at the same time we launch the server. We aim for our first release (deliverable D5) to be in Month 9 of the project using existing gene-PMID links assembled in WP1 together with the sample of gene-PMID links computed using text2seq from WP2 that are available at the time.

Integration with Ensembl database is inherent in our schema, however we will also provide integration with the Ensembl Genome Browser by establishing a DAS server to deliver pubmed2ensembl data in a browsable format. The DAS protocol is specifically designed to provide custom gene and genome annotation data [3], such as that served by pubmed2ensembl. Provision of a pubmed2ensembl DAS source will allow pubmed2ensembl data to be directly incorporated into the Ensembl Genome Browser, allowing users the extremely powerful function of navigating from contig or gene views directly to PubMed records. We will in the first instance use the DAS 1.5 server implementation that is packaged with BioMart, which further justifies the use of BioMart for the rapid deployment of pubmed2ensembl. We will explore the use of alternative DAS servers such as proserver or dazzle if necessary [16]. Although the Ensembl gene DAS reports currently provide limited links to curated references in UniProt, proving that this strategy will work, we note that currently no dedicated resource serves links to the biomedical literature in the DAS registry [17]. pubmed2ensembl data will be served as both gene and region DAS sources. For curated and gene name recognition sources, PMIDs will be linked to all transcripts of a gene and the union of all transcripts for the genomic region. Data from text2seq mappings will be given explicit genomic intervals based directly on their genome mappings as well as associated to the nearest gene.

We will also develop server-side applications that connect pubmed2ensembl to PubMed *via* the NCBI's e-utils as well as to established text mining services provided by NaCTeM, such as Termine or Acromine [18, 19]. These applications will not be visible to the user but will form the basis of the pubmed2ensembl interfaces that will allow users to perform gene or genome specific PubMed queries described in WP4. To minimise the overhead of maintaining a local version of PubMed, to maximise speed, and maintain an up-to-date service, text-based queries will be executed remotely at NCBI using e-utils. Results of remote queries will be cross-referenced locally on the pubmed2ensembl server to find the intersection of PMIDs that are present in gene sets or genomic regions specified by the user.

Work Package 4: Development of pubmed2ensembl human and programmatic web interfaces.

While BioMart and DAS provide excellent off-the-shelf resources for the rapid and successful initial deployment of the pubmed2ensembl system, they are not designed to allow cross-lingual queries over the textual and genomic data. Therefore, we will develop custom interfaces to allow both human and programmatic access to query the pubmed2ensembl database, with a special emphasis on textual queries not supported by current functionality in BioMart. Given the time constraints for the project, we will develop a basic web GUI (deliverable D6), and establish a fully functional API and web services (deliverable D7) that can be programmatically invoked and/or be integrated into bioinformatics workflow management systems, such as Taverna. Development of pubmed2ensembl interfaces will be designed around **Use Cases** (see above) to focus programming effort, provide key deliverables that can aid bench scientists, bioinformaticians and text miners, and to demonstrate the utility of the pubmed2ensembl system.

Statement of Timeliness and Promise

There is an undeniably pressing need for the integration of biomedical text and genomic data, as these are the primary data sources for a large number of biological researchers. Currently, no system exists to provide full integration of text and genomic data, while the development of PosMed and Entrez Gene demonstrates the critical emerging value of such systems. While the engineering of fully integrated portals for text and genome data will require substantial funding beyond the scope of this project, investment in the pubmed2ensembl system will allow the timely production of the key data type (high-quality gene-PMID relations) that will catalyse and form the basis of all such systems in the future.

The potential impact of the pubmed2ensembl system is extremely high since this work will improve access to information in international genome and publication databases, which are used by tens of thousands of biomedical researchers. The scope of the proposed resource is wide, with benefits ranging from bioinformaticians to text miners to bench-biologists, and may provide a key to facilitating the functional analysis of genomes. Both PIs anticipate using pubmed2ensembl resources in future research projects, including: 1) hybrid gene name recognition systems that use both text and sequence information, 2) automated methods to extract Gene Ontology terms from biomedical documents for user-defined sets of genes, and 3) automated reconstruction of regulatory networks based on direct transcription factor – target gene relationships extracted from text and sequence data. We anticipate that the pubmed2ensembl resource will also catalyse research in a wide array of other experimental and computational domains.

Long Term Maintenance of the Resource

Once designed and deployed, the system would require only light maintenance, as imports of external data and in-house mappings between the literature and genome data would be fully automated. The pubmed2ensembl database and server would be maintained as part of core computing facilities supported by the Faculty of Life Sciences, which is supported by three full time computer officers. Updates to text2seq and web interfaces will be maintained by the PIs, while external components of the system will be supported by their respective developers. Further collaborations will be initiated with UKPMC, Ensembl and NaCTeM and additional research project funding will be pursued to further development of the system.

Data Sharing and Dissemination of Results

We will develop and distribute all data and software as open-access or open-source to maximise the use and re-use of the pubmed2ensembl system and its components. Use of pubmed2ensembl data developed in house will be provided with no restriction, and only open-access data from external sources will be included for redistribution. Specifically, access to pubmed2ensembl data will be provided openly through open-access/open-source human and programmatic interfaces, released as an open-access DAS data source, the text2seq application will be distributed as open-source software, and all publications will be submitted to high profile open-access journals and manuscripts will be deposited in UKPMC. pubmed2ensembl will be reported on at ISMB, the main international meeting for Bioinformatics and Text Mining, and linked to from the NaCTeM and Manchester Bioinformatics websites to gain a wide user base. Internal and external tutorials and demonstrations will be conducted where possible. Press releases will be distributed where possible and additional advertising of the resource will take place through various open-access Bioinformatics resource portals.

Training Potential of the Project

The PDRA will be able to take advantage of training opportunities generally offered by the Faculty and specifically by the project. The project will allow training in multi-disciplinary, post-genomic skills including genome bioinformatics and text mining. Research staff in FLS are given an annual performance and development review, and have access to a training programme supported by a full time Research Staff Development Officer. This programme includes monthly training bulletins, one-to-one advice and guidance and workshops on topics such as fellowships, grant reviewing and academic CV writing.

References

1. Hirschman, L., M. Colosimo, A. Morgan, and A. Yeh (2005) *BMC Bioinformatics*. **6 Suppl 1**: S11.
2. Flicek, P., B.L. Aken, K. Beal, B. Ballester, et al. (2007) *Nucleic Acids Res*.
3. Dowell, R.D., R.M. Jokerst, A. Day, S.R. Eddy, et al. (2001) *BMC Bioinformatics*. **2**: 7.
4. Kasprzyk, A., D. Keefe, D. Smedley, D. London, et al. (2004) *Genome Res*. **14**: 160-9.
5. *Entrez Programming Utilities*.
6. Tu, Q., H. Tang, and D. Ding (2004) *Bioinformatics*. **20**: 75-7.
7. Maglott, D., J. Ostell, K.D. Pruitt, and T. Tatusova (2007) *Nucleic Acids Res*. **35**: D26-31.
8. Hoffmann, R. and A. Valencia (2004) *Nat Genet*. **36**: 664.
9. Glenisson, P., B. Coessens, S. Van Vooren, J. Mathys, et al. (2004) *Genome Biol*. **5**: R43.
10. Wren, J.D., W.H. Hildebrand, S. Chandrasekaran, and U. Melcher (2005) *Bioinformatics*. **21**: 4046-53.
11. Fernandez, J.M., R. Hoffmann, and A. Valencia (2007) *Nucleic Acids Res*. **35**: W21-6.
12. Hoffmann, R. and A. Valencia (2005) *Bioinformatics*. **21 Suppl 2**: ii252-8.
13. Jenssen, T.K., A. Laegreid, J. Komorowski, and E. Hovig (2001) *Nat Genet*. **28**: 21-8.
14. Ning, Z., A.J. Cox, and J.C. Mullikin (2001) *Genome Res*. **11**: 1725-9.
15. Kent, W.J. (2002) *Genome Res*. **12**: 656-64.
16. Finn, R.D., J.W. Stalker, D.K. Jackson, E. Kulesha, et al. (2007) *Bioinformatics*. **23**: 1568-70.
17. Prlic, A., T.A. Down, E. Kulesha, R.D. Finn, et al. (2007) *BMC Bioinformatics*. **8**: 333.
18. S. Ananiadou and H. Mima (2000) *International Journal of Digital Libraries* **3**:117-132.
19. Okazaki, N. and S. Ananiadou (2006) *Bioinformatics*. **22**: 3089-95.