Commercialisation Processes in Bioinformatics: Analysis of Bioinformatics Patents

Bruce Rasmussen

*Deputy Director at the Centre for Strategic Economic Studies, Victoria University Melbourne.

CSES Working Paper No. 26

December 2005

Centre for Strategic Economic Studies
Victoria University
PO Box 14428
Melbourne VIC 8001 Australia
Telephone +613 9919 1340
Fax +613 9919 1350
Website: http://www.cfses.com
Email: csesinfo@vu.edu.au

Contact: bruce.rasmussen@vu.edu.au

COMMERCIALISATION PROCESSES IN BIOINFORMATICS: ANALYSIS OF BIOINFORMATICS PATENTS

Bruce Rasmussen
December 2005

Summary

This paper sets out to document the converging nature of the innovation process in bioinformatics; to examine patenting strategies of bioinformatics companies; to identify the nature of the key companies and other organisations in the bioinformatics innovation process; and to identify the key research teams involved. It is very much work in progress.

Given the complex nature of patent analysis and issues with the definition of bioinformatics, much of the paper is devoted to providing an outline of various patent analysis methodologies and the implications of those methodologies for the analysis and conclusions. The key conclusions are summarised below.

Documenting the Reality of Convergence. The analysis of patent subclasses documents key aspects of the process of convergence for the technologies involved in bioinformatics. It shows the expected patent cross classification between IT and biotechnology classes, together with the complex weaving together of information technology and biotechnology to provide new solutions.

Declining Role for Specialist Bioinformatics Companies. The analysis demonstrates the degree to which bioinformatic companies are increasingly focussing on drug discovery and biotechnology more generally. Only 19% of the patents issued by bioinformatics companies between 2000 and 2004 were bioinformatics patents, with the remainder in other areas of biotechnology. For such companies there was a declining trend for bioinformatics patents while, after some allowance for the lag in patent issue, a rising trend for non-bioinformatic patents.

Concentration in a Small Number of Firms. Most of the bioinformatic patents have been issued to a relatively small number of companies and universities. More than half the bioinformatics patents have been issued to 27 organisations and within that group six organisations account for 24% of total patents. These six organisations represent a cross section of the types of organisations classified as part of this analysis – one specialist bioinformatics company (Rosetta Inphamatics), two drug discovery companies (Incyte and CuraGen), one university (Uni of California) and a non biomedical company (IBM).

Geographic Concentration. The concentration of these companies in California is most striking. Of the patents issued to the top 27 assignees, 60% were assigned to Californian companies and universities.

Role of Key Scientific Teams. Most of the leading patent assignees owe their position to less than a dozen highly effective teams of scientists, who were by and large the

'inventors of bioinformatics'. A number of the leaders of these teams went on to found major bioinformatics or related biotech companies.

Introduction

The process of commercialisation of innovation is complex, and this is particularly true of the commercialisation arising from innovations in converging technologies such as biotechnology and information technology. The process of innovation has many systemic features which incorporate networks and feedback loops in which many actors interact in complex ways (see for instance Powell et al. 2005).

Patents and alliances provide documentary evidence of aspects of this process. This paper reports on work in progress on a study of the innovation and commercialisation process of bioinformatics using the US PTO patent database and the Recap database of alliances. The objective of the study is to:

- document the converging nature of the innovation process in bioinformatics;
- examine patenting strategies of bioinformatics companies;
- identify the nature of the key companies and other organisations in the bioinformatics innovation process;
- identify the key research teams involved; and
- examine trends in bioinformatics alliances and the types of the companies involved.

Analysis of patents is difficult. The meaning of trends is not as it appears (Griliches 1990) and the classification systems are not particularly well constructed for sectoral analysis. Nonetheless the patent contains much valuable information including the names of the inventor(s), the original assignee (although not necessarily the current one), the year of filing and year of issue, in addition to details of the invention. The granting of a patent is an important marker in the innovation process and possession of a patent can be essential to the IP position of a technology company. The patent analysis in this paper is based the US PTO's data base which is arguably the world's most important repository of patents, especially in the field of bioinformatics which is dominated by US companies.

There is a range of methodologies used to analyse patents (see for instance Rickne 2000). The essential problem arising from analysing patents relating to a particular technology such as bioinformatics is to devise a classification methodology that identifies the relevant patent and only those patents. The suitability of the methodology depends on the purpose of the analysis. The methodology may be based on the patent sub classes in which the patent is classified. Alternatively it can be based on the patents assigned to identified class of companies eg bioinformatics companies. Thirdly it can be based on the information contained in the patent itself. While it is not practical to review every patent, a common methodology is to undertake key word searches of the patent title, abstract or claim.

However no single methodology is ideal for identifying patents for an emerging technology such as bioinformatics. By its nature it is at the intersection of a number of patent classes. No single word or even several words are sufficient to identity such patents. Accordingly this paper employs three different methodologies to identify bioinformatics patents, each to achieve a different purpose. One of these employs

several methodologies simultaneously to establish a single database of 'bioinformatics' patents.

Outline of Patent Methodologies Used

The first is to use a definition of bioinformatics patents based on several patent subclasses in which bioinformatics patents are expected to be clustered. This has the inherent problem of both including non bioinformatics patents along with bioinformatics patents and of potentially missing bioinformatics patents from other classes. Nonetheless as will be demonstrated the methodology represents a good starting point for examining the cross classification of bioinformatics patents. This analysis illustrates the converging nature of the technologies comprising bioinformatics.

The second approach is to examine patents issued to identify bioinformatics companies as assignee. This shifts the problem to one of defining a comprehensive list of bioinformatics companies. The methodology adopted here is to use a listing of companies involved in bioinformatics alliances from the Recap database. This approach enables the characteristics of patents held by these companies, such as their classification structures, to be explored. As a result it is possible to identify the technologies being patented by the bioinformatics companies and to compare these with technologies revealed by the patent classifications identified in the first approach.

The third methodology adopts a multiple approach. It adopts the first approach, US patent sub classes regarded as 'bioinformatic' to establish an initial set of patents. A multiple word search is undertaken of these patents' titles and abstracts to select out a set of bioinformatics patents. These were then reviewed manually to establish a final list of bioinformatics patents. This patent database has been used to conduct the more detailed analysis of bioinformatics patents such as inventor and assignee.

Issues with patenting bioinformatics innovations

Intellectual property associated with bioinformatics has many dimensions such as lines of code, algorithms, data content and structures and user interfaces. As summarised in the table below taken from Harrison (2003) there are several ways in which bioinformatics IP can potentially be protected – copyright, patent and trade secret. Of these patent protection is of greatest assistance in protecting most forms of IP associated with bioinformatics. This includes lines of code and algorithms that relate to an application, data structure and the user interface. Data content is not protectable by patent.

Copyright offers a way of protecting lines of code and the user interface but tends to be 'thin because it protects unauthorised copying, modification or distribution, not independent development' (Fernandez and Achiriloaie 2004, p. 33) and the trade secret route may help in protecting data content not protectable by patent. Accordingly bioinformatics companies are likely to pursue complex and comprehensive IP protection strategies. Nonetheless seeking patent protection is of particular importance.

Intellectual Property Associated with Bioinformatics Programmes

Component of programme	Types of protection	Notes
Lines of code	Copyright	Protectable: copyright protects the lines of code (both source code and object code)
	Patent	Not protectable: patent protection is excluded for the lines of code. However, the technical idea embodied in the lines of code may be protected in as far as it relates to the technical application of the computer programme
	Trade secret	Source code might be protectable as a trade secret
Algorithm	Copyright	Not protectable: an algorithm cannot be protected by copyright law
	Patent	Not protectable: an algorithm per se is not protectable by patent. The application of an algorithm to a technical application is, however, protectable
	Trade secret	An unpublished algorithm known to only a limited number of people is protectable as a trade secret
Data content	Patent	Not protectable
	Copyright	Not protectable in most countries: compilations of data are protectable in some countries (e.g. Australia
	Database rights	Protectable (in the European Union and some other countries)
	Trade secret	Unlikely to be protectable: except if knowledge of the content of the data is limited to a small number of people
Data structure	Patent	A novel data structure is protectable
	Copyright	The structure of the data is not protectable under copyright. However, drawings showing how the structure is implemented and the lines of code encoding the implementation of the data structure are protectable
	Trade secret	May be protectable if knowledge of the data structure is limited to a small number of people
User interface	Patent	A user interface can be protected by a patent
	Copyright	The 'look and feel' of the user interface can be protected by copyright
	Trade secret	Not protectable, since the user interface would generally be seen by an unlimited number of people
	Design	The aesthetic features of the interface are protectable

Source: Harrison 2003.

The task of successfully patenting bioinformatics innovations, which 'involve applications of computer implemented protocols or software in collecting and/or processing biological data' (Hultquist et al. 2002, p. 743), is challenging. An additional problem is the time for a patent to be issued, 2-4 years during which time innovations seeking patent protection may have become outdated (Fernandez and Achiriloaie 2004).

Bioinformatics innovations fall within the general category of computer related inventions of which there are two aspects – hardware and software. While hardware has always been patentable in the US, patent protection for software has been the subject of intense debate. Initially software was not patentable. However this position was effectively modified in 1996 when the USPTO issued new guidelines for computer related inventions. Of particular relevance to methods of biological data collection and processing, the Guidelines distinguished between data structures, which continued not to be patentable, while a computer readable medium encoded with a data structure was patentable. This was on the basis that the encoded medium 'defines

structural and functional relationships between the data structure and the medium which permit the data structure's functionality to be realised' (USPTO Guidelines quoted in Hultquist 2002, p. 743).

Accordingly the USPTO has issued patent approvals for computer readable media encoded with a computer program for processing and analysing biological data. The Guidelines require such processes to involve pre computer activity or post computer activity. This is generally the case for bioinformatics patent claims of new methods or processes, which generally involve a practical application such as capturing 3-D images of protein structures or sequencing DNA.

The USPTO has responded to the challenge of examining bioinformatics patent applications by establishing in 1999 a new Art Unit (1631) to deal with such claims, developing a collection of bioinformatics patents by searching prior art in likely classes and providing specialised training for examiners in the field (Woodward, n.d.).

Part 1: Using Patents to Document the Converging Nature of the Innovation Process in Bioinformatics

Bioinformatics patent classes in the US Patent Classification System

The USPTO divides the entire set of US patents into searchable collections based on the technology claimed in accordance with the US Patent Classification System. The primary groupings are known as *classes* based on technology associated with a particular industry or subject matter having a similar function, use or structure. Classes are subdivided into small ordered collections of patents called *subclasses*, which are the smallest searchable collection. Related subclasses are nested together under a *mainline* subclass with those subclasses listed lower in the schedule being more specific than those above (USPTO 2004). For instance, the class containing many of the bioinformatics patents is 702 Data Processing: Measuring, Calibrating or Testing. The subclasses of particular relevance to bioinformatics are 702/19 to 702/21. These fall within the *mainline* subclass 702/1 Measurement System in a Specific Environment. Of the subclasses 702/19 to 702/21, 702/19 'biological or biochemical' is the most general of the three subclasses. Subclass 702/20 is concerned with 'gene sequence and determination' while 702/21 relates to 'cell count or shape or size analysis'.

While Woodward (n.d.) alluded to the high likelihood of bioinformatics patents being classified within the range 702/19 to 702/32 when the newly established Art Unit 1631 sought to identify prior bioinformatics patents, other studies have sought to be both broader and more specific in their selection of subclasses relevant to bioinformatics. This reflects the desire to capture patents reflecting the convergence of the two underlying technologies, while not being so broad as to include patents of no relevance (see for instance Patel 2003; Gatto 2001).

As its initial definition of bioinformatics patents, this study adopts that defined in Patel (2003), namely subclasses 702/19 to 702/21, 703/11, 703/12 and 382/129. Class 703 is Data Processing: Structural Design, Modelling, Simulation and Emulation. Subclasses 703/11 and 703/12 relate to biological, biochemical and chemical aspects of mainline subclass 703/6, Simulating non-electrical devices or systems. Subclass

382/129 falls in the general class 382 Image Analysis and relates in particular to 'DNA or RNA pattern reading'. These subclasses are all a priori relevant to various aspects of bioinformatics, i.e. data processing (classes 702 and 703) that includes measuring, calibrating, testing structural design, modelling and simulation relating to biological, biochemical and chemical functions and uses. The inclusion of subclass 382/129 appears justified on the basis that the application of image analysis to DNA or RNA pattern reading is a major means of bioinformatics data generation. The adopted subclasses are listed in the table below.

US Patent Sub Classes Used to Define Bioinformatics Patents

US PTO sub class no.	US PTO sub class description
382/129	Image Analysis and relates in particular to 'DNA or RNA pattern reading'
702/19	Data Processing: Measuring, Calibrating or Testing – Biological or biochemical
702/20	As for 702/19 but relating specifically to 'Gene sequence determination'
702/21	As for 702/19 but relating specifically to 'cell countor shape or size analysis'
703/11	Data Processing: Simulating Non electrical Device or System – Biological or biochemical
703/12	Data Processing: Simulating Non electrical Device or System – Chemical

Typically patents are classified as belonging to more than one class reflecting the multifaceted aspect the technologies involved in most patent applications. This is likely to be particularly the case for bioinformatics patents, which involve technologies that are found, not only in the data processing patent classes, but also in the biotechnology patent classes. For instance a bioinformatics patent may be classified not only in subclass 702/19, but also to one or more biotechnology subclasses in class 435 Chemistry: Molecular Biology and Microbiology. It is also likely to be classified in one or more of the closely associated subclasses in 702 involving various aspects of data processing. Many patents classified in 702/19 are also classified in one or more of the other 702 subclasses, such as 702/27, which is concerned with molecular structure or composition. In this way it is possible to examine the technologies involved in each patent issued.

The classification structure of bioinformatics patents

The table below records the number of subclasses listed for each bioinformatics patent issued for 2004 using the Patel 2003 definition. There were a total of 163 bioinformatics patents issued by the USPTO in 2004, based on this definition, and on average each patent was classified into 5.7 patent subclasses. The class entries for each of the bioinformatics subclasses have been cross tabulated to illustrate the extensive linkages between the bioinformatics patent subclasses and other related classes and subclasses. The table shows the sum of the patent subclass for each defined bioinformatics subclass. For instance, there were a total of 22 patents classified to 382/29. Of these patents 17 were classified to 702/19 and 2 to 702/20. None were classified to either 703/11 or 703/12. However in total these 22 patents were also classified 79 times to other 382 subclasses and eighteen times to 702 classes other than 702/19-21. Of particular significance to the convergence of IT and biotechnology is the cross classification with the biotech subclasses. The 22 382/19 patents are cross-classified 82 times to subclasses in the major biotech class 435.

This analysis allows three categories of linkages to be observed.

Patent Subclass Frequency: Bioinformatics Patents Issued by the USPTO in 2004

	382/29	702/19	702/20	702/21	703/11	703/12
Bioinformatics subclasses						
382/129	22	17	2	1	0	0
702/19	17	112	22	1	8	4
702/20	2	22	41	1	0	0
702/21	1	1	1	6	0	0
703/11	0	8	0	0	18	8
703/12	0	4	0	0	8	21
Related non-bioinformatics subclasses						
Other 382	79	70	4	10	5	5
Other 702	18	88	13	7	9	18
Other 703	0	12	0	0	21	27
Other 700	1	24	7	7	8	26
Biotech subclasses						
435	82	166	54	2	11	0
436	1	31	9	1	3	3
530	0	18	9	0	0	0
536	14	31	13	0	0	0
Total biotech	97	246	85	3	14	3
All other	10	90	21	1	14	24

Source: USPTO CSES analysis.

The first is the level of cross classification between the six bioinformatics classes. This is strongest between 382/19 (imaging) and 702/19 (biological or biochemical data processing) with 17 out of 22 or 77% of 382/19 patents also classified as 702/19. Other strong cross classifications are between closely related subclasses. Fifty four per cent or 22 of 41 of 702/20 patents (a specialised gene sequencing subclass of 702/19) are also classified as 702/19. The two 703[1] subclasses concerned with measuring, calibrating or testing, 703/11 (biochemical) and 703/12(chemical) have quite high levels of cross classification. However there are few links between either702/21 (cell count and size analysis) and 703/12 and the other bioinformatics classes.

The second set of cross linkages illustrated by the table are between the bioinformatics subclasses and the larger biotechnology classes 435 (Chemistry: Molecular Biology And Microbiology), 436 (Chemistry: Analytical And Immunological Testing) and the more specialised class 536 (Carbohydrates or derivatives e.g., pectin, glycosides, RNA, DNA, cellulose, starch, etc.). To measure the strength of these linkages, the number of cross linked subclasses has been divided by the relevant number of patents in each bioinformatics subclass. For instance on average each 382/19 patent is also classified in 3.7 435 subclasses. The measure is crude in the sense that Class 435 contains more subclasses than Class 536 and would be expected to 'score' higher on that basis than 536. Nonetheless comparisons between the bioinformatics codes for each biotech class and the total for all biotech classes is an indicator of the complexity of the technology being patented and the

[1] Class 703 is Data Processing: Structural Design, Modeling, Simulation and Emulation.

level of its cross linkages with other biotech subclasses. These measures are set out in the table below.

Average number of jointly classified subclasses for each bioinformatics subclass

	382/29	702/19	702/20	702/21	703/11	703/12
Bioinformatics subclasses						
382/129	**1.0**	0.2	0.0	0.2	0.0	0.0
702/19	0.8	**1.0**	0.5	0.2	0.4	0.2
702/20	0.1	0.2	**1.0**	0.2	0.0	0.0
702/21	0.0	0.0	0.0	**1.0**	0.0	0.0
703/11	0.0	0.1	0.0	0.0	**1.0**	0.4
703/12	0.0	0.0	0.0	0.0	0.4	**1.0**
Related non-bioinformatics subclasses						
Other 382	3.6	0.6	0.1	1.7	0.3	0.2
Other 702	0.8	0.8	0.3	1.2	0.5	0.9
Other 703	0.0	0.1	0.0	0.0	1.2	1.3
Other 700	0.0	0.2	0.2	1.2	0.4	1.2
Biotech subclasses						
435	3.7	1.5	1.3	0.3	0.6	0.0
436	0.0	0.3	0.2	0.2	0.2	0.1
530	0.0	0.2	0.2	0.0	0.0	0.0
536	0.6	0.3	0.3	0.0	0.0	0.0
Total biotech	**4.4**	**2.2**	**2.1**	**0.5**	**0.8**	**0.1**
All other	0.5	0.8	0.5	0.2	0.8	1.1

The strongest links are between subclasses, 382/129, 702/19, 702/20 and the biotech classes. On average each 382/129 patent is also classified in biotech subclasses 4.4 times and for 702/19 and 702/20 the average is about 2. This confirms expectations that bioinformatics patents would have strong cross linkages to biotechnology. The linkages with the other bioinformatics subclasses 702/21, 703/11 and 703/12 is much lower. The strongest is with 703/11 with on average 0.8 biotech subclasses. However the linkages with the other two subclasses, 702/21 (0.5) and 703/12 (0.6) is comparatively so low as to place doubt on the appropriateness of including them in the definition of bioinformatics subclasses.

The third set of cross linkages are between the defined bioinformatics subclasses and related non bioinformatics classes such as other subclasses in the same class or similar classes. For instance each 382/19 patent is also classified on average in other 382 subclasses 3.6 times. Similarly 702/19 patents are reasonably frequently cross classified with other 702 subclasses (average 0.8) as well having relatively high linkages with non-bioinformatics subclasses in 382 (average 0.6). The other subclasses, 703/11 and 703/12 exhibit high cross linkages with other 703 subclasses with averages of 1.2 and 1.3 subclasses respectively. Comparisons between these averages need to be adjusted for the number of subclasses in each class. For instance Class 382 has 225 subclasses and Class 702 has 199 subclasses while Class 703 has only 28. On this basis 703/11 and 703/12 are strongly connected to other subclasses in their class.

In general these cross linkages reflect the complexity of the technologies being patented. Not only are there the cross linkages between ICT subclasses and biotech but also within ICT and biotech subclasses. However these cross linkages between related subclasses may also be an indicator of subclasses that should be included in the definition of bioinformatics patents and those that could be excluded. This analysis suggests for instance that the key bioinformatics subclasses are 382/129, 702/19 and 702/20 with strong cross classification to biotech subclasses. Subclasses 702/21, 703/11 and 703/12 seem to be less relevant. On the other hand sub class 702/27 (Data Processing: Measurement System - Molecular structure or composition determination) was frequently cross classified with bioinformatics subclasses, particularly 702/19 and may on this basis be a candidate for inclusion as a bioinformatics subclass.

Part 2: An Analysis of the Patenting Strategies of Bioinformatics Companies

Bioinformatics Company Patents

An alternative approach to exploring bioinformatics patents is to examine the classification structure of patents issued to identified bioinformatics companies.

Patent Structure: Selected Bioinformatics Companies 2004

	Bioinformatics	Non bioinformatics	Total
Bioinformatics subclasses			
382/129	2	0	2
702/19	19	0	19
702/20	8	0	8
702/21	1	0	1
703/11	5	0	5
703/12	1	0	1
Related non-bioinformatics subclasses			
Other 382	6	1	7
Other 702	14	3	17
Other 703	4	3	7
Other 700	8	22	30
Biotech subclasses			
435	32	189	221
436	6	14	20
514	2	24	26
530	0	16	16
536	9	137	146
546	0	38	38
Other biotech	0	21	21
Total Biotech	49	439	488
All other	10	34	44
Total No. of subclasses	127	502	629
No. of patents issued	25	106	131

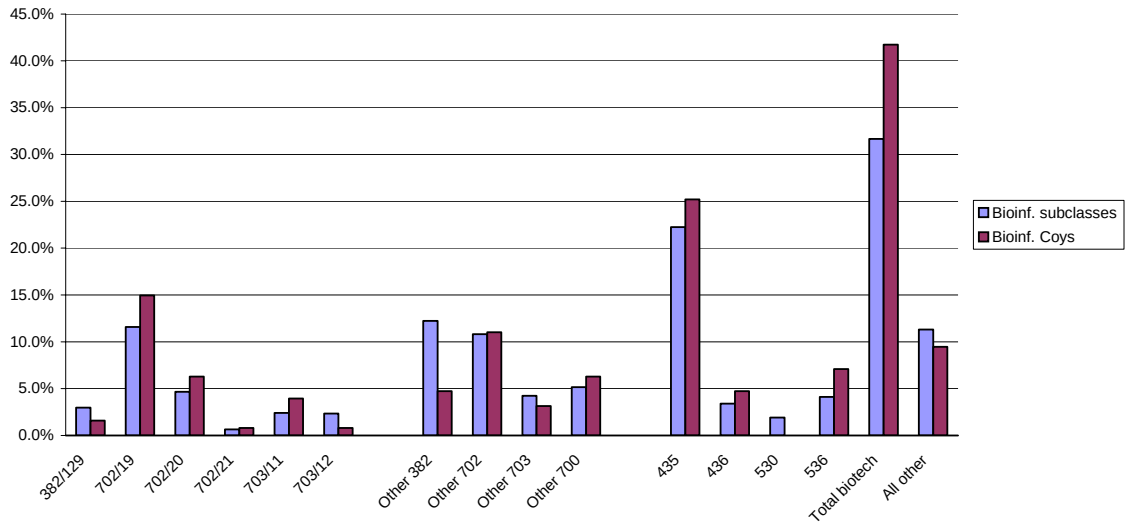
In this way it is possible to examine the nature of the IP being patented by bioinformatics companies as distinct from patents classified as 'bioinformatics'. A group of 20 bioinformatics companies was selected from earlier work on biotechs that was most prominent in establishing bioinformatics alliances. The group of 20 are selected from those with the largest number of bioinformatics alliances listed on Recap for the period 1990 to 2005.

Details of the patents issued to these companies as assignee by the USPTO for the period 2000 to 2004 were downloaded from the USPTO site and their classification structure analysed. The period selected covers the large majority of bioinformatics patents issued. The table above summaries the results of this analysis.

The Patel definition appears to be reasonably robust in distinguishing between bioinformatics and non bioinformatics patents. In the section above it was suggested that the inclusion of 702/27 and the exclusion of 703/12 might improve the definition. For this group of patents the inclusion of 702/27 would appear to be advantageous. It would transfer one patent currently misclassified as a non-bioinformatics patent to bioinformatics. The exclusion of 703/12 would have no impact, as the one patent affected would remain as a bioinformatics patent by virtue of subclass 703/11.

Further evidence is provided by the similarity in subclass structure for the two sets of bioinformatics patents as shown in Figure 1 below. Those selected from the group of bioinformatics companies have a slightly higher proportion of subclasses in 702/19 and in the biotech subclasses. Neither of which suggests they are not bioinformatics patents. The proportion of 382/129 and other 328 subclasses in the bioinformatics company patents is significantly lower than for the patents selected by the bioinformatics definitions. These subclasses are concerned with image analysis and tend to be issued to non bioinformatics companies.

Figure 1. Bioinformatics Patent Classification Structure: Comparison of Bioinformatics Companies and All Companies

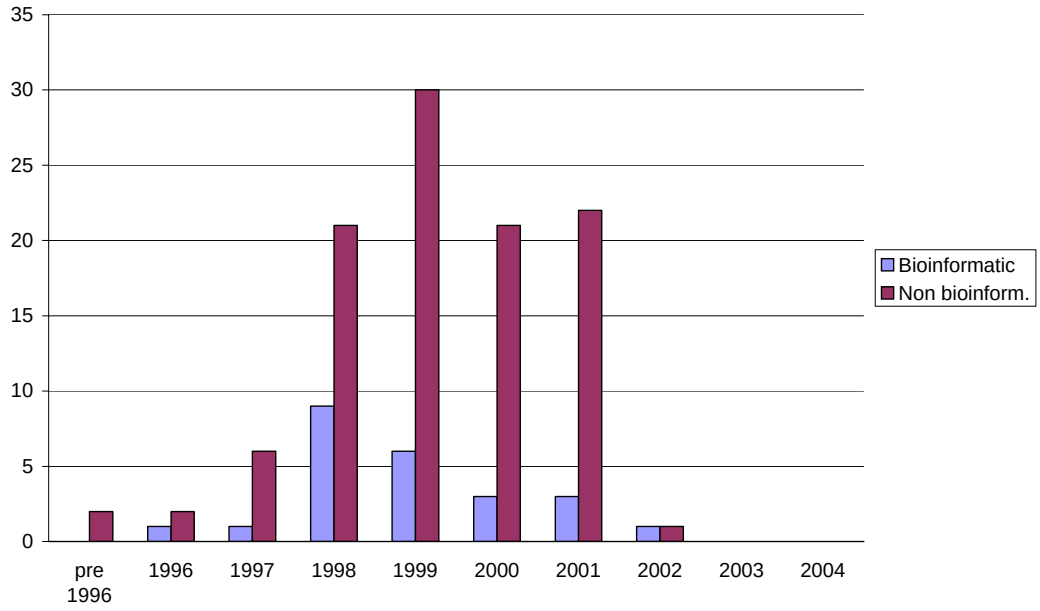


If this result is correct, it means that the vast majority of patents assigned to bioinformatics companies are not bioinformatics patents. As indicated by the subclasses to which they are classified the remaining patents are biotechnology patents. This may reflect two factors. One is the relative difficulty of patenting

bioinformatics innovations. The other may reflect a shift in the strategy of bioinformatics companies towards drug discovery and away from generic bioinformatics services.

Figure 2 shows a comparison of bioinformatics and non-bioinformatics patents by file date for patents issued to bioinformatics companies over the period 2000-04. This indicates that while the trend in non-bioinformatics patents has been relatively steady since 1998, filings for bioinformatics patents has been in decline.

Figure 2. No of Patents by File Date for Patents Issued to Selected Bioinformatics Coys 2000-04



There is a lag of several years between filing and successfully obtaining a patent (issue date), which explains why the most recent filing shown is for 2002. The average lag for this group of patents is 3 years and is about six months longer for bioinformatics than non-bioinformatics patents. However this is not sufficient to explain the strongly divergent trends in patent filings.

This analysis supports the proposition that bioinformatics companies have been switching their patenting activity from bioinformatics to drug discovery and other biotechnology-focussed innovations.

Part 3: Using Patents to Identify the Key Players (Inventors Companies and Other Organisations) in the Bioinformatics Innovation Process

Some prior studies of the innovation process of new technologies has highlighted the prominent role of a relatively small number of inventors, companies and other organisations in the discovery and commercialisation of such technologies (see for instance Braun and MacDonald 1978). To test this proposition for bioinformatics we sought to identify the assignees and inventors with multiple holdings of bioinformatics patents. In order to do this a database of bioinformatics patents was

extracted from the US PTO, which could be more closely interrogated. For this database to be meaningful it was necessary, while building on the earlier work, to more closely define bioinformatics patents.

Further Classification of Bioinformatics Patents

In Part 1, the Patel definition of bioinformatics patents, namely those patents classified as belonging to US Patent Sub classes 382/129, 702/19, 702/20, 702/21, 703/11 and 703/12 were examined in some detail. This work suggested that subclasses 702/19 and 382/129 were the core bioinformatics subclasses. The inclusion of subclasses 702/20 and 703/11 appeared to usefully include additional bioinformatics patents, while the analysis suggested that this may not be the case for 702/21 and 703/12 which might contain a majority of non-bioinformatics patents. Accordingly while such a definition may be satisfactory as an indicator of broad trends, it has the potential to be misleading when used as a basis for a more detailed examination of the commercialisation of bioinformatics.

The classification process used here is to adopt a three stage process to establish a database of bioinformatics patents which can be examined in more detail. Firstly all patents falling into the 'Patel' subclasses were downloaded from the USPTO database for the period 1997 to April 2005. There were very few patents in these classes granted prior to that period. Following elimination of duplicates this produced a database 725 patents.

Secondly a multiple key word search of the patent title and abstract was undertaken of this group of patents. Two sets of key words were used reflecting the converging nature of the technologies as identified in Part 1. One set was of key words that reflected the biotechnology nature and the other the information technology content of bioinformatics. To qualify a patent needed to satisfy both tests. This produced a list of 552 patents.

Thirdly all patents were reviewed manually to eliminate any areas of doubt and a final decision made, in some cases, by reviewing the patent claims and other patent details. In forming an opinion about the inclusion of particular patents reference was made to the following US PTO definition of a bioinformatics patent. This definition suggests that bioinformatics can be defined as:

- 'the use of computational methods to solve biological problems;
- involves storage, retrieval and analysis of biological data; and
- will aid in gene discovery, molecular modelling, mutagenesis.

and involves methods reciting significant data processing steps directed to inventions in areas such as gene sequencing, nucleic acid hybridization, protein structure prediction, X-ray crystallography, pharmacology, receptor-ligand modelling, and immunology' (Woodward, n.d.). This process resulted in 365 patents remaining in the database.

While this definition excluded patents that clearly had no relevance to biotechnology, it also excluded some patents with applications to health services such as patient diagnostics. To have extended the definition to the health IT domain would have involved more subclasses than those considered here and gone beyond the generally accepted definition of bioinformatics which is grounded in the need for data management solutions in biological research (see for instance Maojo and Kulkowski 2003).

Analysis of Assignees

Types of assignees
This analysis has been conducted on a narrower group of 332 patents for which the assignee has been classified according to type. For the assignees included in this group, from bioinformatics companies such as Rosetta, to large universities such as the University of California, the importance of bioinformatics ranges from the essential to the incidental.
Biotechnology companies form the largest group of bioinformatics related companies. These include:

- specialist bioinformatics companies such as Rosetta, LION Bioscience for which bioinformatics is central to their business activities;
- other general platform technology companies such as Affymetrix, that use their expertise in bioinformatics to combine with other platform technologies to provide a range of technology products or services. For such companies the bioinformatics component is the software that manages the functioning of an instrument and processes any data output; and
- drug discovery companies for which bioinformatics is but one of a number of platform technologies that helps guide their drug discovery activities.

Other companies involved in bioinformatics include pharmaceutical companies, as well as a range of non medical companies, such as IBM Life Sciences, for which bioinformatics forms an important part of its business. Universities and research institutes are also important in developing bioinformatics IP and are significant holders of bioinformatics patents.

Accordingly for this detailed analysis, assignees were classified, using annual reports and web based company information, into six organisational groups:

Biotechnology companies comprising:
Bioinformatics companies
Drug discovery companies
General platform technology (GPT) companies
Non biomedical companies
Pharmaceutical companies
Universities/research institutes

The table below shows the number of organisations in each category and the number of patents assigned to each type of organisation. The table shows the range of organisations involved in the bioinformatics innovation process. Universities have a prominent role. The largest number of assignees is universities 45 or 31% and they have 81 or 24% of the patents. There are only 11 bioinformatics companies out of a total of 146 assignees with 36 or 11% of the patents. However they have the highest average number of patents 3.3 compared with 2.3 for all organisational categories. Other biotechs such as those involved in drug discovery or GPT number 21 and 35 respectively with 16% and 33% of the patents respectively. Altogether the biotechs represent 46% of assignees and 60% of patents. Indicating the relative importance of bioinfoamtics patents for these companies. Non biomedical companies are the third largest category with 25, but hold only 13% of the patents.

Patents held by Type of Assignee: 1997 to April 2005

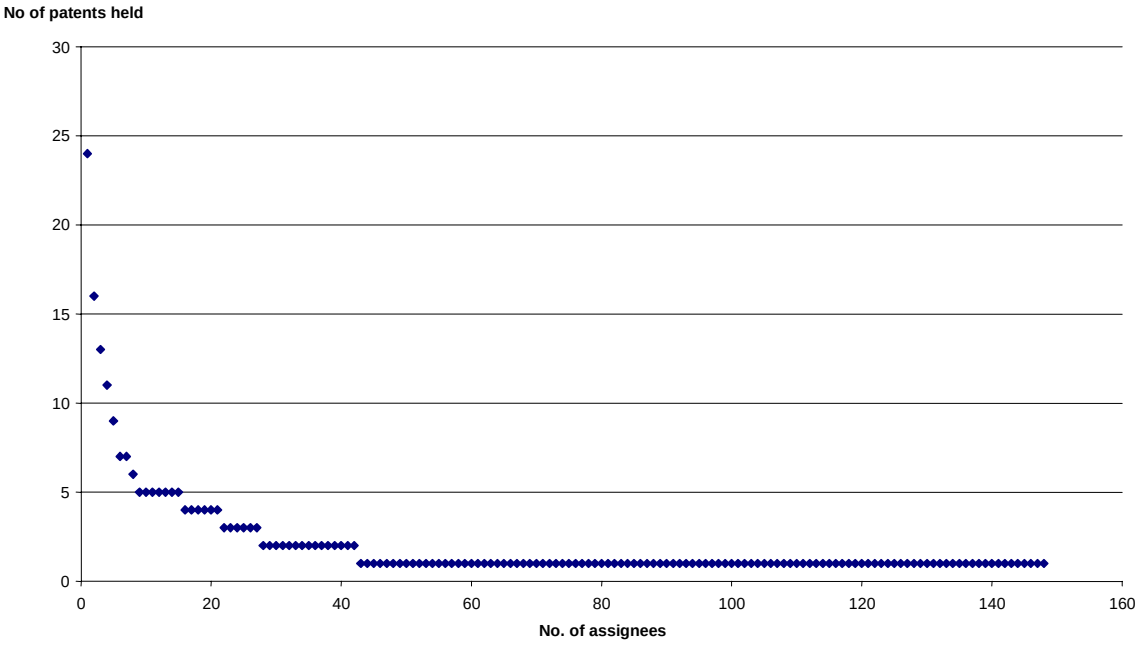
Assignee type	No. of org.	% org. type	No. of patents	% of patents	Av. no. of patents
Bioinformatics	11	8%	36	11%	3.3
Drug Discovery	21	14%	52	16%	2.5
GPT	35	24%	110	33%	3.1
Non_Biomedical	25	17%	42	13%	1.7
Pharma	9	6%	11	3%	1.2
Uni	45	31%	81	24%	1.8
Total	146	100%	332	100%	2.3

Note: Excludes patents with no or not classifiable assignee.

Concentration of Patents in small number of assignees

While the table above is instructive in providing an indication of the spread of the types of organisations involved in the bioinformatics innovation process it sheds little light on the major players involved. As hypothesised earlier Figure 3 shows that bioinformatics patents have been assigned to a relatively small group of assignees. A group of only 27 organisations have three or more patents each and together have been assigned more than half of the patents in this database.

Figure 3. Bioinformatics Patents Issued 1997 to April 2005



A further analysis of these organisations is shown in the table below, the largest number of patents, 24, is held by Affymetrix, a leading GPT company. The second ranked is the University of California with 16 and Rosetta Imphamatics, a bioinformatics company, is next ranked with 13. Two drug delivery companies Incyte and CuraGen together have 20 patents. The largest non-biomedical company holding, IBM is next with 7 patents. Together these top six organisations are assignees for 80 patents or 24% of the total. It is also noteworthy that these six organisations cover each of the organisation types representing the dominant positions in each category.

Top Assignee Organisations by No. of Patents Assigned

Assignee:	Bioinf	DD	GPT	Non_Bio medical	Uni	Grand Total
Affymetrix Inc. (Santa Clara, CA)			24			24
The Regents of the University of California (Oakland,CA)					16	16
Rosetta Impharmatics, Inc. (Kirkland, WA)	13					13
Incyte Corporation (Palo Alto, CA)		11				11
CuraGen Corporation (New Haven, CT)		9				9
International Business Machines Corporation (Armonk, NY)				7		7
Agilent Technologies Inc. (Palo Alto, CA)			7			7
California Institute of Technology (Pasadena, CA)					6	6
Tripos, Inc. (St. Louis, MO)	5					5
The John Hopkins University School of Medicine (Baltimore, MD)					5	5
Large Scale Biology Corporation (Vacaville, CA)			5			5
Applera Corporation (Foster City, CA)			5			5
Hitachi, Ltd. (Tokyo, JP)				5		5
Cytokinetics, Inc. (South San Francisco, CA)		5				5
BioDiscovery, Inc. (Los Angeles, CA)	5					5
Visible Genetics Inc. (Toronto, CA)			4			4
ViaLogy Corp. (Altadena, CA)	4					4
University of Utah Research Foundation (Salt Lake City, UT)					4	4
The Scripps Research Institute (La Jolla, CA)					4	4
Oxford GlycoSciences (UK) Ltd (Abingdon, GB)			4			4
Fujitsu Limited (Kawasaki, JP)				4		4
Vertex Pharmaceuticals Inc. (Cambridge, MA)		3				3
TissueInformatics, Inc. (Pittsburgh, PA)			3			3
Stanford University (Palo Alto, CA)					3	3
Sangamo Biosciences, Inc. (Richmond, CA)*		3				3
Pharmacopeia, Inc. (Cranbury, NJ)		3				3
Northeastern University (Boston, MA)					3	3
Total	27	34	52	16	41	170

Key
Biotech
Uni
Non Biomedical
* Also involved in drug discovery.

A majority (16) of the companies in the top 27 are biotechs, with a particular interest in bioinformatics– either bioinformatics specialists, GPT or drug discovery companies. There are six universities and only three non-biomedical companies. One other striking feature of the table is the predominance of Californian companies and universities. While fourteen of the top 27 assignees are Californian, 102 of the total of 170 patents are held by Californian organisations. This is indicative of a strong bioinformatics cluster in California.

The inventors

Many of the organisations listed in the table above owe their prominent position as assignees in bioinformatics patents to the innovative teams of scientists that they employed to generate these patents.

The table above sets out the leading inventors of bioinformatics patents (3 or more) together with their assignee organisations. There is a close match between the leading inventors and the leading assignees listed in the table above.

Number of Bioinformatics Patents by Assignee and Inventor 1997 to 2005

Assignee	Inventor	Number
Affymetrix Inc	Balaban David	7
Affymetrix Inc	Berno Anthony J.	3
Affymetrix Inc	Hubbell Earl	3
Affymetrix Inc	Lipshutz Robert J	4
Affymetrix Inc	Mittmann Michael P	3
Affymetrix Inc	Walker Michael G.	3
Affymetrix Inc	Webster Teresa A.	3
Agilent Technologies Inc.	Wolber Paul K.	3
BioDiscovery Inc	Shams Soheil	4
California Institute of Technology	Dahiyat Bassil I.	5
California Institute of Technology	Gordon D. Benjamin	5
California Institute of Technology	Street Arthur	5
California Institute of Technology	Mayo Stephen L	5
CuraGen Corporation	Deem Michael W	8
CuraGen Corporation	Rothberg Jonathan Marc	9
CuraGen Corporation	Simpson John W.	8
Cytokinetics Inc.	Adams Cynthia L.	4
Cytokinetics Inc.	Crompton Anne M.	4
Cytokinetics Inc.	Vaisberg Eugeni A.	5
Cytokinetics Inc.	Sabry James H.	4
Incyte Corporation	Akerblom Ingrid E	5
Incyte Corporation	Au-Young Janice	4
Incyte Corporation	Cheng Rachel J.	3
Incyte Corporation	Goold Richard D	3
Incyte Corporation	Hibbert Harold H.	4
Incyte Corporation	Hillman Jennifer L.	4
Incyte Corporation	Klingler Tod M.	5
Incyte Corporation	Maslyn Timothy J.	3
Incyte Corporation	Walker Michael G.	3
Large Scale Biology Corporation	Anderson N. Leigh	4
Large Scale Biology Corporation	Anderson Norman G	4
Large Scale Biology Corporation	Goodman Jack	3
Northeastern University	Karger Barry L.	3
Northeastern University	Miller Arthur W.	3
Oxford GlycoSciences	Amess Robert	3
Oxford GlycoSciences	Bruce James Alexander	4
Oxford GlycoSciences	Platt Albert Edward	3
Oxford GlycoSciences	Prime Sally Barbara	3
Oxford GlycoSciences	Stoney Richard Michael	3
Oxford GlycoSciences	Parekh Rajesh Bhikhu	4
Rosetta Impharmatics Inc	Bassett, Jr. Douglas	3
Rosetta Impharmatics Inc	Bondarenko Andrey	3
Rosetta Impharmatics Inc	Friend Stephen H.	5
Rosetta Impharmatics Inc	Stoughton Roland	8
Sangamo Biosciences Inc	Jamieson Andrew	3
Sangamo Biosciences Inc	Rebar Edward J.	3
Sangamo Biosciences Inc	Eisenberg Stephen P.	3
The Regents of the University of California	Marcotte; Edward M	4
The Regents of the University of California	Eisenberg; David	5
Tripos, Inc	Patterson; David E	3
ViaLogy Corp	Gulati; Sandeep	4

Each of the major companies had teams of scientists who took out sizeable patent portfolios – many sharing as co- inventors on the patent.. The leading scientists in some of the teams had 5 or more patents. These are listed below and may be regarded as representing the key inventors of the bioinformatics technology.

Leading Bioinformatics Inventors by No. of Patents 1997 to April 2005

Organisation	Inventor	No.
Affymetrix Inc	Balaban David	7
California Institute of Technology	Dahiyat Bassil I.	5
California Institute of Technology	Gordon D. Benjamin	5
California Institute of Technology	Street Arthur	5
California Institute of Technology	Mayo Stephen L	5
CuraGen Corporation	Deem Michael W	8
CuraGen Corporation	Rothberg Jonathan Marc	9
CuraGen Corporation	Simpson John W.	8
Cytokinetics Inc.	Vaisberg Eugeni A.	5
Incyte Corporation	Akerblom Ingrid E	5
Incyte Corporation	Klingler Tod M.	5
Rosetta Impharmatics Inc	Friend Stephen H.	5
Rosetta Impharmatics Inc	Stoughton Roland	8
The Regents of the University of California	Eisenberg; David	5

Those scientists from the private sector placed their companies in highly privileged positions with respect to the commercialisation of bioinformatics technology and personally played a key role at those organisations. For instance, Stephen Friend currently heads up Rosetta, now a division of Merck and Jonathan Rothberg, until recently CEO and President of CuraGen.

Conclusion

This study has set out to:

- document the converging nature of the innovation process in bioinformatics;
- examine patenting strategies of bioinformatics companies;
- identify the nature of the key companies and other organisations in the bioinformatics innovation process;
- identify the key research teams involved; and
- examine trends in bioinformatics alliances and the types of the companies involved.

Given the nature of the paper, as a report on work in progress, it has also provided an outline of various patent analysis methodologies and the implications of those methodologies for the analysis and conclusions.

The analysis of patent subclasses has provided an indication of the complex process of innovation for converging technologies such as bioinformatics. Not only is there the expected patent cross classification between IT and biotechnology classes, but also evidence of the complex nature of the information technology and biotechnology itself being woven together to provide new solutions.

The examination of the patents issued to bioinformatics companies demonstrates the degree to which these companies are increasingly focussing on drug discovery and biotechnology more generally. Only 19% of their patents issued between 2000 and

2004 have been bioinformatics patents with the remainder being in other areas of biotechnology. There was a declining trend for bioinformatics patents while that of non-bioinformatics, after some allowance for the lag in patent issue, has been rising.

The study of assignees and investors has been instructive about the process of innovation. Most of the patents have been issued to a relatively small number of companies and universities. More than half the bioinformatics patents have been issued to 27 organisations and within that group six account for 24% of total patents. These six organisations represent a cross section of the types of organisations classified as part of this analysis – one specialist bioinformatics company (Rosetta Inphamatics), two drug discovery companies (Incyte and CuraGen), one university (Uni of California) and a non biomedical company (IBM). The concentration of these companies in California is most striking. Of the patents issued to the top 27 assignees, 60% were assigned to Californian companies and universities.

It is evident from the analysis of inventors that most of the leading patent assignees owe their position to less than a dozen highly effective teams of scientists, who were by and large the' inventors of bioinformatics'. A number of the leaders of these teams went on to found major bioinformatics or related biotech companies.

This research continues to raise many questions. For instance the work relating patents to alliances is still being undertaken but is expected to provide further insight into the dynamics of the innovation and commercialisation process. More work is required on the performance of firms that gained prominent positions in bioinformatics IP. More broadly we need a better understanding of how the innovation and commercialisation processes identified here compares with that of other converging platform technologies.

References

Braun, E. and MacDonald, S. 1978, *Revolution in Miniature: The History and Impact of Semiconductor Electronics*, Cambridge University Press, Cambridge, Mass.

Fernandez, D and Achiriloaie, M. 2004, Keeping-up intellectual property lifelines for life science ventures', *Journal of High Technology Law*, vol. 3, no. 1, pp. 29-39.

Gatto, J. 2001, Bioinformatics Patents: Challenges and Opportunities', *Bioinformatics Advisory*, November, Mint, Levin, Cohn, Ferris, Glovsky and Popeo PC.

Griliches, Z. 1990, 'Patent statistics and economic indicators: A survey', *Journal of Economic Literature*, vol. 28, pp. 1661-1707.

Harrison, R. 2003, 'Protecting innovation in bioinformatics and in-silico biology', *Biodrugs*, vol. 17, no. 4, pp. 227-231.

Hultquist, S.J., Harrison, R. and Yang, Y.Z. 2002, 'Patenting bioinformatic inventions: Emerging trends in the United States', *Nature Biotechnology*, vol. 20, no. 7, pp. 743-744.

aojo, V. and Kulikowski, C.A. 2003, 'Bioinformatics and Medical Informatics: Collaborations on the Road to Genomic Medicine?', *Journal of the Americakn Medical Informatics Association*, vol. 10, no. 6, pp. 515-522.

Patel, P. 2003, 'UK Performance in Biotechnology-related Innovation: An Analysis of Patent Data', Final Report prepared for the Assessment Unit of the UK Department of Trade and Industry, SPRU, University of Sussex, Brighton.

Powell, W., Koput, K., White, D. and Owen-Smith, J. 2005, 'Network Dynamics and Field Evolution: The Growth of Interorganizational collaboration in the life sciences', *American Journal of Sociology*, vol. 110, no. 4, pp. 1132-1205.

Rickne, A. 2000, New Technology-based Firms and Industrial Dynamics: Evidence from the Technological System of Biomaterials in Sweden, Ohio and Massachusetts', Doctoral Thesis, Chalmers University of Technology, Goteborg, Sweden.

USPTO 2004, 'Overview', US Patent and Trademark Office, Washington DC, available at www.pto.gov

Woodward, M. n.d. 'Smippets and Bytes', Powerpoint presentation, US Patent and Trademark Office, Washington DC, available at www.uspto.gov/web/patents/biochempharm/documents/**snippets**.pps