Copenhagen
Business School
HANDELSHØJSKOLEN

Biotech Business Working Paper No. 04-2004

Heterogeneity of Intellectual Assets

A Method for Identification and Measurement With Patent Data

By Henrich Dahlgren, Rasmus Lund Jensen & Finn Valentin

research centre on
biotech business

RESEARCH CENTRE ON BIOTECH BUSINESS ¨ COPENHAGEN BUSINESS SCHOOL ¨ SOLBJERGVEJ 3, 3 ¨ DK-2000 FREDERIKSBERG ¨ DENMARK
TEL.: +45 3815 2505 ¨ FAX: +45 3815 2540 ¨ WEBPAGE: WWW.CBS.DK/BIOTECH ¨ E-MAIL: BIOTECH@CBS.DK

Heterogeneity of Intellectual Assets

A Method for Identification and Measurement With Patent Data

By Henrich Dahlgren*, Rasmus Lund Jensen & Finn Valentin

Research Centre on Biotech Business
Copenhagen Business School
Solbjergvej 3,3
DK – 2000 Frederiksberg

Tel: +45 3815 2560
Fax: +45 3815 2540
hd.ivs@cbs.dk

*) Corresponding author

Abstract

This paper deals with methodological issues of measuring and assessing the composition and level of heterogeneity of firms' intellectual assets. It develops an original metric - referred to as the H-index - for measuring heterogeneity at firm level using data extracted from patent documents. The main purpose is to improve the characterization of research activities within firms in the biotechnology sector. Although the H-index grew out of research on biotech firms, the metric carries broader relevance for all patent-intensive industries. The measurement and assessment of the H-index is illustrated and tested using empirical data from our study on Scandinavian biotech firms.

ISBN 87-91549-03-5

2nd edition

Introduction

This paper addresses methodological issues of measuring and assessing the composition and level of heterogeneity of firms' intellectual assets. It presents an original metric for measuring heterogeneity, using data extracted from patent documents. We developed this metric - referred to as the *H-index* - with the purpose of improving the characterization of firms in the biotechnology sector. The core of this sector, Dedicated Biotech Firms (DBFs), has the production of drug discovery research as its sole economic activity, undertaken primarily on the basis of intellectual assets. This type of output is inadequately characterized by standard classifications of industries or products, rendering public statistics of limited use. However, outputs from biopharmaceutical research are intensely protected by IPR, making patent documents an attractive alternative data-source. Although the H-index grew out of research on biotech firms, the metric carries broader relevance for all patent-intensive industries.

Terminologically speaking, the concept of "heterogeneity" in this context is preferred above its semantic cousin, "diversification". The latter emphasizes the extension of assets or activities from a given point of departure, and is often associated with an analytical focus on the strategy and direction of the change by which this extension comes about. In comparison, heterogeneity is a more straightforward phenomenon, essentially concerned with the degree of dissimilarity between constituent elements of a composite configuration, assuming no particular locus or level of original homogeneity.

The paper starts with a brief overview of the theoretical issues that have spurred an interest in measuring knowledge heterogeneity. Second, a brief review of the literature on diversification and heterogeneity reveals similarities and differences between various methods of measuring inter-firm and intra-firm heterogeneity, and we argue there is a need for an additional method. Third, we present and discuss the H-index in more detail. In this section, we describe the process of measuring intra-firm heterogeneity by transforming standard patents codes (IPC) into a corresponding classification system, referred to as *H-codes*. Finally, we discuss the assessment of the H-index and corresponding composition and level of heterogeneity and test the validity of the metric using empirical data from our study on Scandinavian biotech firms.

The theoretical context

A variety of theoretical issues in the literature on industrial dynamics have stimulated development of different methodologies to measure heterogeneity and diversification.

Diversification, and its effects on firm performance, has formed the basis for a key research question in the strategic management literature. Since performance differences, as an outcome of related and unrelated diversification, has been a major issue of concern in this research, a number of methodologies have been developed to measure the degree of heterogeneity between different parts of diversified firms (Ramanujam and Varadarajan 1989; Chatterjee and Wernerfelt 1991; Markides and Williamson 1994; Breschi *et al.* 2003).

Increasing diversification of the knowledge and technology base of companies accentuates the strategic issue of managing diversity. Theoretically, three key forces are argued to drive technological diversification within firms; (a) opportunities to improve products and production systems through the implementation of new technologies; (b) continuing competitive advantage of existing technologies, and finally (c) changes in products, production systems, and supply chains caused by innovation. Technological diversification is, in turn, argued to be a driving force behind firm growth, increasing R&D investment, and the emergence of new business opportunities based on related technologies (Granstrand *et al.* 1997).

Firms tend to operate on the basis of an increasing number of technologies (Granstrand 1998) and Granstrand *et al.* (1997) found that firms tend to diversify into a higher number of technological fields than product classes. Diversification occurs, however, primarily within related technological fields (Patel and Pavitt 1997).

From another perspective, complementarities may cause firms to diversify into different technological fields when, for instance, they wish to diversify their range of related products (Pavitt 1998). This is especially seen in large firms, although their technological profiles seem to be rather stable over time (Patel and Pavitt 1997).

The concept of path dependency is central to the evolutionary theory of firms and industries (Nelson and Winter 1982; Metcalfe and Gibbons 1989). Understanding path dependency in the formation of competencies translates into issues of identifying continuity and cumulativeness in learning and innovation in companies. Relatedness and similarity – the inverse of heterogeneity – become central issues in this strain of literature, as exemplified by (Miyazaki 1995; Miyazaki 1999). Bibliometric and patent-based data are used to measure the level of cumulativeness.

Data and metrics

Before presenting indicators of heterogeneity based on patent data, we briefly discuss other public data sources on industries, products, and labour markets, leaving aside, however, methods and indicators that require questionnaires or case studies.

Industry classification codes (ICCs)

Studies involving measurement of the heterogeneity of diversified firms have used Industry Classifications Codes (ICCs), such as NACE and SIC, as their primary source of data. However, for several reasons the use of ICCs is not entirely unproblematic for this purpose.

First, the level of heterogeneity of firms materialises in ICCs codes only when the company is organised into different business units, and the propensity to do so is not a simple reflection of its heterogeneity. E.g. companies of sufficient size and heterogeneity in which the decomposability (Simon 1996) of assets is aligned with the composition of their outputs will typically pursue an M-form organisation (Chandler 1962) with separate business units, with ICC codes meaningfully indicating their underlying heterogeneity. Companies characterised by non-decomposable bundles of technologies and outputs (Prencipe *et al.* 2003) tend to maintain their U-form, with none of their heterogeneity expressed into different ICC codes.

Second, ICCs are used to categorise firm activities based on their outputs. However, based on the similarity of outputs, firms may be categorised into the same industry, although the heterogeneity of their internal asset compositions may vary considerably. Firms differ in the level of outsourcing, or may produce similar products using different process technologies.

Third, this latter source of distortion is exacerbated in industries undergoing rapid technological transformations, to which firms are known to adapt with different styles and timing. And finally, in emerging technologies, statistical classifications exhibit serious time lags in their responsiveness to the challenge of categorising new types of firms. To exemplify, based on careful scrutiny of the Danish population of biotech firms, the present authors have identified in Denmark a total of 45 genuinely research based firms. The relevant NACE category is "Research and development, natural science and techniques" (code no. 731000). Table 1 shows that more that 25% of Danish DBFs are categorised into other codes than 731000.

Table 1 Distribution of NACE codes on Danish DBF's

NACE	Frequency	%	Description
244100	2	4,4	Production of pharmaceutical products
244200	2	4,4	Pharmaceutical production plants
295690	1	2,2	Manufacturing of production equipment
652395	1	2,2	Financing services
730000	3	6,7	Research and development
731000	32	71,1	Research and development, natural sciences
732000	3	6,7	Research and development, social sciences
743000	1	2,2	Technical testing and analysis
	45	**100**	

Product classification

One alternative to ICC is classification at the product level, of the type applied in SITC (Standard International Trade Classification). Again, the question is to what extent the output (products and services) is a representative or useful indicator of internal asset structures. A given set of intellectual assets may be used for producing several kinds of related products and services. Also, firms sometimes need to manage more technologies than what can be directly derived from observation of the kind of output generated. Hence, the relationship between the intellectual asset structure and the output is not unambiguous.

Input-output flows between industries

Diversification at the firm level has been measured with the use of data on input-output flows between industries. Scherer (1982) measured relatedness at industry level based on R&D expenses in one industry and the use of the generated output in other industries. The 'concentric index' developed by Caves *et al*. (1980) has been used in several subsequent studies, for example by Montgomery and Wernerfelt (1988), and bears some resemblance to the metric introduced in this paper. The concentric index measures diversification by taking into account the percentage of a firm's sales in different industries and a weighted value for the degree of similarity between SIC codes among the industries in which the firm is engaged. The index is calculated as

$$\mathrm{D}_i = \sum_{j=1}^{n} m_{ij} \sum_{l=1}^{n} m_{il} r_{jl}$$

where D_i is the diversification of firm *i*, determined by the percentage of firm *i*'s sales in industry *j* (m_{ij}), the relatedness of other industries *l* in which firm *i* is engaged (r_{jl}) and the percentage of firm *i*'s sales in other industries (m_{il}). The weighted relatedness (r_{jl}) is assigned value *0* if industry *j* and *l* have the same three-digit code, value *1* if industry *j* and *l* have the same two-digit code but different three-digit codes, value *2* if they have different two-digit codes, and value *3* if they have different one-digit codes.

Input-output flows between industries rely on industry classification codes, making them vulnerable to the limitations discussed above.

Educational composition of workforce

In some countries, labour market statistics offer information on the educational attributes of the workforce. Particularly in the Scandinavian countries, the national statistical agencies have developed linkages from this information to statistics using firms as their level of analysis. As a result, firms may be characterised by the educational composition of their employees, and these characteristics may be translated into measures of the heterogeneity of the intellectual assets of firms. Using this methodology, Vejrup-Hansen examined the intellectual assets of engineering consultancy firms in Denmark (Valentin and Vejrup-Hansen 2004), while Oskarsson (1993) studied large Swedish firms in electronic engineering.

Patent classification

A number of studies have used patent classification data to indicate diversification. The International Patent Classification (IPC) system is used by patent examiners for categorizing and organizing patents within different technological fields[1]. Codes distinguish between a large number of technological fields, primarily for search purposes, and have been increasingly used for measuring diversification among firms (e.g. Jaffe 1986; Jaffe 1989; Granstrand *et al.* 1997; Patel and Pavitt 1997; Patel and Pavitt 2000). Firms' technological competencies and diversification into different technological fields are often measured using firms' patent shares in different IPC codes, often in the form of revealed technological advantage. The latter is defined as the relative importance of the firm in each field of technological competence, after normalizing for the firm's share of total patenting. Jaffe (1986) and Jaffe (1989) focused on the relatedness of technologies at the firm level. Jaffe measured technological relatedness between firms using patent classifications by applying the 'cosine index', which has been frequently used to measure the level of similarity of research in technological fields between firms. The similarity, or "technological proximity", is measured by calculating the overlap of the fractions of firm A's and firm B's patents in different patent classes. Technological proximity is, applying the cosine index, given by

[1] For further information about IPC, please refer to www.wipo.int.

$$P_{ij} = \frac{\sum_{k=1}^{K} f_{ik} f_{jk}}{\sqrt{\left(\sum_{k=1}^{K} f_{ik}^{2}\right)}\sqrt{\left(\sum_{k=1}^{K} f_{jk}^{2}\right)}}$$

where f_{ik} and f_{jk} are the fractions of firm *i*'s and firm *j*'s patents, respectively, in patent class *k*. P_{ij} is the degree of overlap between f_i and f_j. P_{ij} is 1 when f_i and f_i are identical, and 0 for firms without any overlap of patents.

Verspagen (1997) and Breschi *et al*. (2003) use the co-occurrence of patent classification codes to describe relatedness between different technological fields. The assumption is that strong relationships between technological fields, measured as the frequency of co-occurrence in patent classifications, may shed light on the degree of "knowledge relatedness" and knowledge spillover between technologies. Breschi *et al*. (2003) apply the cosine index to measure knowledge relatedness and its effect on firms' technological diversification. The studies are based on co-occurrences of IPC codes in patents.

Firms' propensity to patent varies across industries, so clearly there are sectors for which this type of data provides insufficient coverage. Limitations to the validity of patents have also been argued to stem from its inability to pick up the tacit dimensions of the knowledge of firms. Patel and Pavitt (1997) argue, however, that tacit and codified knowledge are complementary, and hence reduces the limitations of patent data as a source for measuring heterogeneity.

Additional uses of patent data

Patent documents offer additional possibilities for characterising the composition of inventor firms. Citations to non-patent literature, i.e. primarily academic papers, are listed so as to document sources and antecedents of the inventions, and they may be translated into information on the composition of firms (Narin 2000). Citations of previous patents may be applied in a similar manner (Granberg 1988).

Patents also have text sections such as titles and abstracts allowing inventions to be characterised systematically. Such characterisation may be extended to the entire patent portfolio of firms, which in turn offers indications of the composition of their knowledge and R&D assets. Characterisation may be based on co-word analysis (e.g. van Raan and Engelsman 1993; Engelsman and van Raan 1994) and its further extension into text-mining methodologies (Valentin and Jensen 2003).

The H-index

The H-index is calculated for single firms, based on the main IPC codes of their patents. To build a metric particularly suited for the technological fields related to the biotech industry IPC codes are translated into an adjusted classification system, referred to as H-codes. Within specific technology fields, IPC codes offer rather fine-grained categories, that many categories are used too infrequently to accommodate statistical purposes. In these cases categories have been combined, based on their technological proximity.

In the translation into 3-level H-Codes, the first level indicates the highest aggregation of technological fields and level 3 the most detailed specification of technological fields. Based on an analysis of IPC codes assigned to the patents held by Danish and Swedish biotech firms, these codes could be re-classified into nine different, homogenous categories at H-code level 1, and further categorized at level 2 and 3 in more detailed sub-categories. IPC code level 1 and 2 have been collapsed into H-code level 1, IPC code level 3 corresponds to H-code level 2, and IPC code level 4 to H-code level 3 (see further figure 1). Use of further levels of the IPC codes would generate too many categories at H-code level 4 and produce an excess level of detail for our study with a skewed number of patents in each subcategory[2].

Figure 1 Translation of IPC codes into H-codes

The five IPC levels of IPC code "C12Q-001/18" denote the following:	
IPC-Level 1:	C: Chemistry and metallurgy
IPC-Level 2:	C12: Biochemistry; beer, spirits, wine or vinegar; microbiology or enzymology, mutation or genetic engineering.
IPC-Level 3:	C12Q: Measuring or testing processes involving enzymes or micro-organisms, compositions or test papers therefore, processes of preparing such compositions, condition-responsive control in microbiological or enzymological processes,
IPC-Level 4:	C12Q-001: Measuring or testing processes involving enzymes or micro-organisms and compositions therefore and/or processes of preparing such compositions.
IPC-Level 5:	C12Q-001/18: Measuring or testing processes involving viable micro-organisms testing for anti-microbial activity of a material.
This IPC code is translated into the H-code "7.2.0" in the following way: Category "7" corresponds to C12 in the IPC system. Category "2" corresponds to Q on IPC level 3. Category "0" corresponds to 001 on IPC level 4.	

Calculating the H-index

The approach of the H-index is similar to Caves' concentric index (Caves *et al.*, 1980) in that it measures weighted dissimilarities between the codes representing different technological fields. H-index is given by

$$H\text{-}index = \frac{\sum_{x=1}^{N-1}\left[\sum_{y=1}^{N-x} r_{P_x P_{x+y}}\right]}{\frac{N(N-1)}{2}}$$

The level of heterogeneity within each firm (*H-index*) is measured as the sum of the weighted relationship values ($r_{P_x P_{xi+y}}$) of patent H-code relationships $\left\lfloor P_x P_{x+y} \right\rfloor$ between all patents (*N*) held by a firm, normalized by the total number of relationships $\frac{N(N-1)}{2}$ between all patents. For each relationship between patents, it is recorded whether H-codes are identical or different. Non-identical relationships are assigned specific *r* for each of the three levels in the H-code. Differences at level 1 are assigned a weight (*r*) of

[2] For further information about the translation of IPC codes into H-codes, please contact corresponding author.

1. Differences at level 2 are given a *r* of *0,5*, and differences at level 3 are assigned a *r* of *0,25*. Relationships between patents with identical H-codes score *0*. Patent relationships are assigned a *r* on one level only. That is, if H-codes differ at level 1, no further evaluation is made at lower level and a value of *1* is assigned to the relationship. If codes are identical at level 1, we check for non-identities at level 2 assigning them the value of *0,5* or proceed to level 3. Firms with none or one patent only are omitted from the calculation and not assigned a H-index value.

To illustrate the calculation of the H-index and to see the effects of the number of patents and the distribution of patents in different technological fields, two cases of H-index calculation are shown in table 2. Patents of Firm A are found in three different technological fields at H-code level 1, indicated by three different codes (3, 5, and 8). Patents in category 3 at H-code level 1 are all found in category 2 at H-code level 2, and in category 3 at H-code level 3. Patents in category 5 at H-code level 1 are found in three different categories at H-code level 2, and so forth. To calculate the H-index, the H-code of patent 1 is compared with H-codes of patent number 2 to 10, the H-code of patent 2 is compared to the H-codes of patent 3 to 10, and so on. For instance, the H-code of patent 1 compared to patent 2 and 3 respectively, scores a *r* of 0 since they are identical. The H-code of patent 1 compared with patent 4 scores a *r* of 1 since they differ at level 1. The H-code of patent 5 compared to that of patent 6 is assigned a *r* of 0,5 since H-codes differ at level 2. The total *r* is divided by the total number of relationships between patents, resulting in a H-index of 0,77. Firm B, on the other hand, hold patents in a single technological field. That is, the H-code is similar at all three levels. Consequently, all patent relationships are assigned *r 0* because H-codes are identical and the H-index of Firm B is *0*.

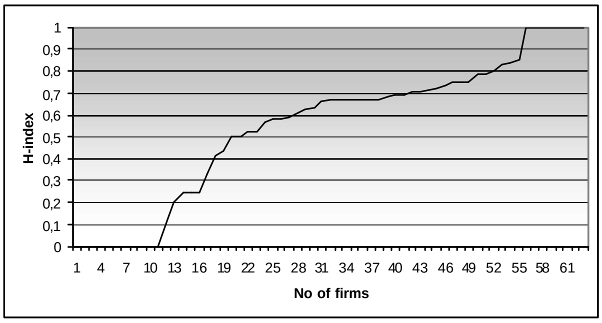
Table 2 H-indexes for a sample of two DBFs

FIRM A	**H-index: 0,77**	**Patents: 10**		
Patent no	Main IPC	H-code	H-code	H-code
1	**A61K-031/045**	**3**	**2**	**3**
2	**A61K-031/167**	**3**	**2**	**3**
3	**A61K-031/167**	**3**	**2**	**3**
4	**C07C-275/00**	**5**	**1**	**5**
5	**C07D-207/00**	**5**	**2**	**0**
6	**C07K-014/705**	**5**	**5**	**3**
7	**G01N-033/48**	**8**	**0**	**2**
8	**G01N-033/50**	**8**	**0**	**2**
9	**G01N-033/50**	**8**	**0**	**2**
10	**G01N-033/68**	**8**	**0**	**2**

FIRM B	**H-index: 0,00**	**Patents: 4**		
Patent no	Main IPC	H-code	H-code	H-code
1	**A61K-031/12**	**3**	**2**	**3**
2	**A61K-031/135**	**3**	**2**	**3**
3	**A61K-031/136**	**3**	**2**	**3**
4	**A61K-031/5375**	**3**	**2**	**3**

Figure 2 exhibits the distribution of H-index on Danish and Swedish DBFs holding more than two patents each.

Figure 2 Distribution of H-index on Danish and Swedish DBFs (N=63)



Comparison of Metrics

The H-index builds on elements also found in Caves' Concentric index. In the following, differences between the two metrics will be examined including comparison with the Herfindahl index. The latter is a widely used metric for measuring heterogeneity and builds on elements similar to the Concentric index.

The Herfindahl index, given by

$$D_i = 1- \sum_{j=1}^{n} m^2$$

and the Concentric index, given by

$$D_i = \sum_{j=1}^{n} m_{ij} \sum_{l=1}^{n} m_{il} r_{jl}$$

where m is the share of a firm's total sales in industries which it is engaged in.

Comparing these two metrics with each other and with the H-index brings out differences in the ways they (1) relate technological categories to each other, (2) assign weights to different levels, and (3) normalize the index. These dissimilarities produce different sensitivity in the three metrics, particularly for firms with few patents, which typically is the case for the population of biotech firms for which the index was originally developed. The following section examines some of the differences.

Relatedness and weights

In its original form, the Herfindahl index assigns no weights for dissimilarities between industry or patent classifications. The Concentric index, on the other hand, takes into account bilateral dissimilarities between all technological fields, applying an ordinary scale of weights (r) ranging from 0 to 2, depending on the level at which dissimilarities occurs.

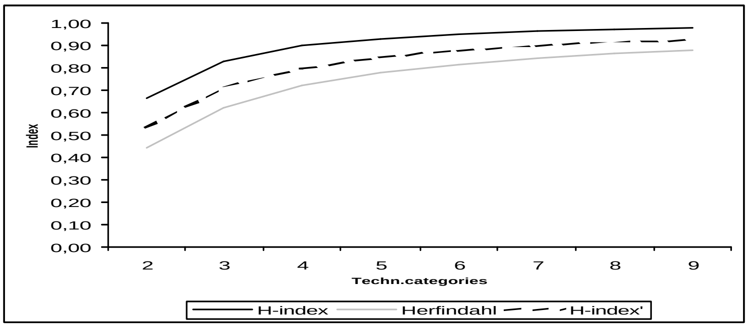
The H-index applies a relative scale of weights (*r*), emphasizing the relative importance of dissimilarities at different levels of H-code. Same 3-digit H-codes is assigned the value 0, different 3-digit H-codes but same 2-digit H-codes is assigned the value 0,25, different 2-digit H-codes but same 1-digit H-codes is assigned the value 0,5, and different 1-digit H-codes scores 1. That is, dissimilarities at higher levels of H-codes indicate wider distance between technological fields and weights are doubled for each higher level of H-codes.

Normalization

All indexes normalize index values, but do so by different approaches. The Herfindahl and Concentric indexes normalize for the relative size of technological categories. Measuring all dyadic patent relationships, the H-index normalizes by the total number of bilateral patent relationships, thus capturing both differences in the number of technological categories *and* the number of patents.

To illustrate the difference, we disregard different weighting and assign dissimilarities between technological categories a score of *1*. Figure 3 shows the effect on index values when a firm increases the number of patents in a given number of technological categories in which it holds patents. When the number of patents increases proportionally, so as to retain the relative share of the total number of patents of each technological category, the Herfindahl and Concentric indexes remain unchanged since the relative sizes of technological categories are constant[3]. This is illustrated by the grey curve. The H-index, on the other hand, shifts downwards illustrated by the black curve and the dotted black curve (H-index'), where the changing index value of a given number of technological categories can be followed along a vertical line. Hence, increasing patenting activity within a given number of technological categories, where the relative size of categories is kept constant, leads to lower H-index. The H-index, in other words, brings out more clearly the size of the actual heterogeneity in the light of total potential heterogeneity with increasing number of patents.

Figure 3 Index comparison



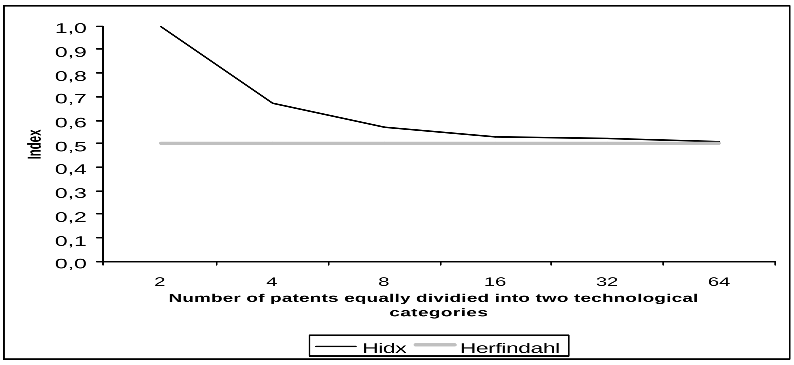
[3] Neutralizing the weights used by the Concentric index by assigning *r* the value of 1 for all differences between technological categories, results in similar index values for the Herfindahl and the Concentric indexes.

Small numbers

The H-index captures differences in both the number of technological categories *and* the number of patents offers by measuring and normalizing by the total number of bilateral patent relationships. Hence, it provides better scrutiny, compared to the other indexes, of small patent portfolios, which we typically find in small biotech firms.

Figure 4 exhibits effects of increasing the number of dissimilar patents in a given number of technological categories, as also shown by the shifting curve in figure 3. When the relative size of technological categories retain the same, the H-index value changes as the number of patents within given categories increases because it is normalized by the total number of bilateral relationships. The other indexes remain constant. The sensitivity of the H-index is higher for small patent portfolios and diminishes as the size of patent portfolios increases.

Figure 4 Index comparison



In one extreme case, the H-index remains constant if a firm holds one patent in each technological category, shown in figure 5. An increasing number of technological categories with one patent in each category will show no effect on the H-index value. The index value is constantly *1*, indicating fully diversified patent portfolios. The other indexes exhibit an increasing heterogeneity as the number of technological categories increases.

Figure 5 Index comparison

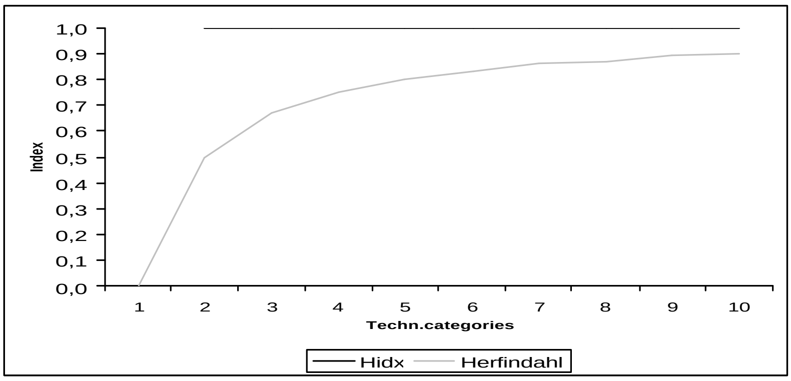
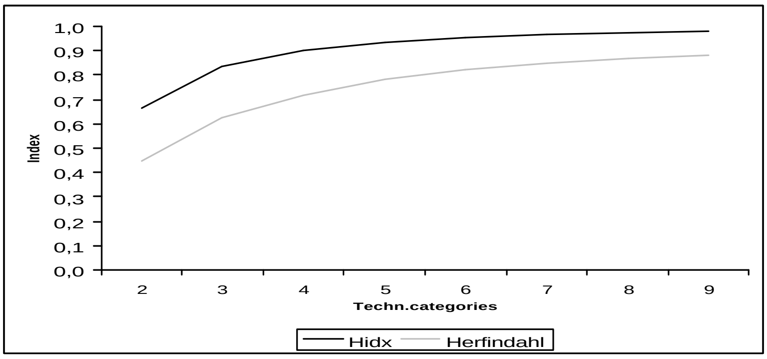


Figure 6 exhibits index values for patent portfolios with a skewed number of patents in the categories, where all categories holds one patents except for one category holding an increasing number of patents. As seen, the indexes exhibit different levels and slopes for small patent portfolios but moves towards similar values for large patent portfolios. For patent portfolios with distribution of patents in technological categories, different from the examples discussed above, the indexes show similar pattern. Hence, different sensitivity for small patent portfolios and moving towards similar index values the larger the patent portfolio.

Figure 6 Index comparison



Application of the H-index

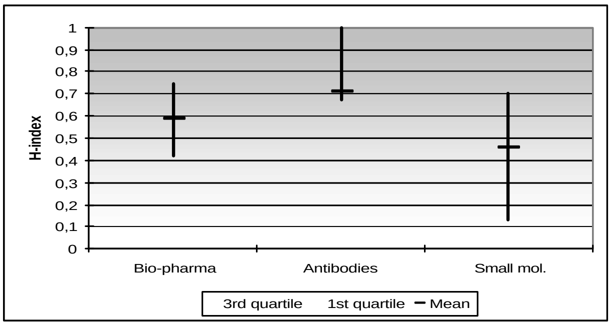
This section exemplifies an application of the H-index on the biotech industry. It is useful to distinguish between the three discovery approaches typically found among dedicated biotech firms: (1) bio-pharmaceuticals (mainly proteins operating as drugs), (2) antibodies (belong to bio-pharmaceuticals but is analyzed separately because of substantial differences in targets and pathways), and (3) small molecules (drugs based on molecules of lower complexity than the former two).

These three approaches differ in the nature of their core problems, and hence also in the composition of the knowledge developed for their solution (Valentin *et al*., 2005). A key issue in *bio-pharmaceuticals* (including antibodies) is to understand the complexity of lead molecules to an extent that permits them to be re-engineered so as to address highly specific targets and pathways for controlled therapeutic effects. In these research approaches, key intellectual assets must combine understandings of both complex pathways and intricate lead molecules. Their highly composite architectures expectedly should be reflected in high H-index values.

Small molecule drug discovery operates with leads of much lower complexity, but face the challenge of achieving complicated therapeutic effects by means of chemical design. For that reason, small molecule firms often focus on a specific binding site and the parts of therapeutic pathways that are immediately connected to that site, which in turn may relate to multiple disease groups. E.g. some receptor families are present in the membranes of many different cell-types in the human body. Correctly understood and approached, these receptors may open up to pathways relevant for multiple potential decease targets, and in small molecule approaches these potentials are more readily explored through the very high variability in synthetic compounds that may be generated and screened. For these reasons, small molecule firms build knowledge that is more focused on specific binding sites and their potential pathways. These efforts will typically be expressed in broader variability in the leads they explore and patent. Their knowledge heterogeneity therefore may be expected to be smaller and more directly associated with variability in target exploration.

Differences between the three discovery approaches are presented in figure 7 which gives mean and inter-quartiles of the H-index in each research strategy, confirming that bio-pharmaceuticals and antibodies show higher mean H-index and their distributions are concentrated at higher levels of H-index.

Figure 7 Distribution of H-index on research approaches indicated by the 1st and 3rd quartile in combination with the mean value (vertical line)



To understand these differences we need a few general concepts on the types of knowledge assets that DBFs bring to bear on their discovery processes. The range of competencies in DBFs involved in drug discovery may be classified in three main types of intellectual assets (Valentin *et al*. 2005):

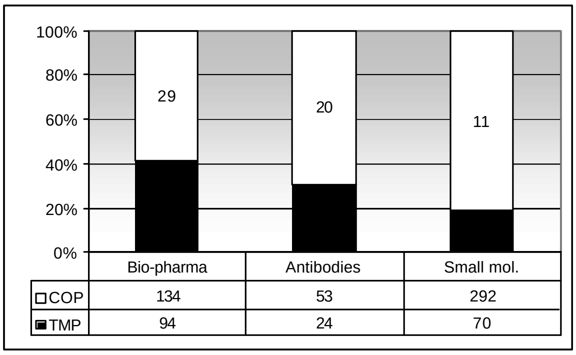
- *Conceptual frameworks* include theories, models and heuristics specifying or suggesting causal relationships and the conditions under which they are operative.
- *Methods* include tools, procedures and research instrumentation for generating, processing and interpreting data. High-throughput screening is probably the most well-known data processing tool for effective screening of high amounts of data
- *Internally generated information* such as screening libraries or other results of previous transformations of data into higher-order inputs for problem solving.

The differences between the three discovery approaches also mean that they vary substantially in the ways they draw on and combine the three asset types, and in turn that should produce different values on the H-Index.

To examine differences in the way these knowledge assets are combined in discovery activities we apply a simplified categorization of patents based on their relationships to the three knowledge assets, as identified by their main IPC. Patents referring directly to compounds and to protein leads, we argue, relate particularly to the conceptual framework assets of DBF, because they are embodied expressions of the way targets and pathways are perceived and modelled. These compound patents will be referred to as COP. The remaining patents of firms, not relating to specific compounds or leads, will instead represent the other two types of assets, i.e. methods and information, and they are referred to as tools/method patents (TMP).

Bio-pharmaceuticals, combining complex knowledge of both pathways for controlled therapeutic effects and re-engineered lead molecules addressing specific targets, is expected to combine a COP share with a comparatively larger share of TMP patents than small molecule drug discovery. The latter discovery approach, on the other hand, is expected to show higher shares of COP patents since they focuses on leads of much lower complexity and primarily strives to achieve complicated therapeutic effects by means of chemical design. The share of COP and TMP patents in firms pursuing each of the three discovery approaches is presented in figure 8.

Figure 8 Share TMP and COP patents in research approaches
(number of firms indicated inside and number of patents beneath columns)

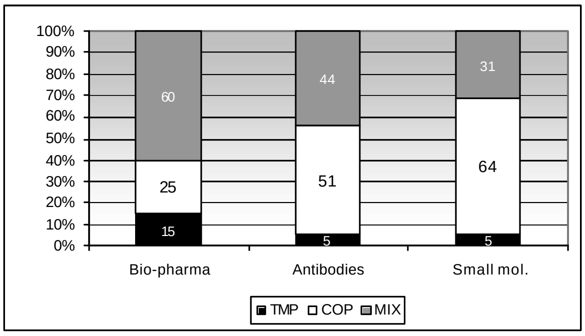


	Bio-pharma	Antibodies	Small mol.
□COP	134	53	292
■TMP	94	24	70

Antibodies and other bio-pharmaceuticals are seen in figure 8 to tend towards more equal shares of COP and TMP patents, whereas small molecule firms are more one-sided, having predominantly one type of patent. The latter research approach, accordingly, have *less composite* knowledge structures than do the two former.

To assess the impact of COP and TMP patents on the H-index, we decompose the overall *r* for each firm into three separate components; a *r* associated with H-code differences *within* COP patents (*r*COP), a *r* associated with H-code differences *within* TMP patents (*r*TMP), and a *r* associated with H-code differences *between* COP and TMP patents (*r*MIX). The sum of these three components corresponds to the total *r* as obtained in the H-index calculation. Each sum is divided by the aggregate sum, revealing their share of the total *r*. For each firm the share of total *r* for *r*COP is calculated, along with corresponding shares for *r*TMP and *r*MIX. The average shares for firms in each research approach are presented in figure 9.

Figure 9 Share of total *r* for patent categories in each research approach.



The total *r* for biopharmaceutical firms has a very high share of *r*MIX (60%), while the second highest share (44%) is observed for antibody firms. Small molecule firms in particular reveal high impact of *r*COP (64%). Figure 9 confirms that each of the components of *r*COP, *r*TMP AND *r*MIX has a share of overall *r* corresponding to what we should expect from the composition of COP and TMP patents presented in figure 8. The expectations are based on the various compositions of knowledge and their relation to core problems in different discovery approaches as discussed in the beginning of this section. In other words, the H-index proves capable of picking up these core differences in knowledge heterogeneity between large and small molecule drug discovery.

Final Remarks

The H-index seems useful, compared to the Herfindahl and Concentric indexes, to measure and assess the level of heterogeneity based on how it relates technological fields, assigns weights to different levels, and how it normalizes index values. The resulting sensitivity is particularly suitable for small numbers of patents, which is shown by the above application of the H-index to biotech discovery approaches. The H-index may, however, be used for analysis of firms in all patent-intensive sectors. In a parallel paper the authors further demonstrate its usefulness for unpacking and analysing architectures of knowledge assets and their related scope advantages in biotech firms (Valentin *et al*. 2005).

References

Breschi, S., F. Lissoni and F. Malerba (2003) 'Knowledge-relatedness in firm technological diversification', Research Policy, 32(1), 69-87.

Caves, R. E., M. E. Porter A. M. Spence (1980), Competition in the Open Economy: A Model Applied to Canada, Cambridge, MA.: Harvard University Press.

Chandler, Alfred D. (1962), Strategy and structure -chapters in the history of the industrial enterprise, Cambridge Mass.: M.I.T. Press.

Chatterjee, S. and B. Wernerfelt (1991) 'The Link between Resources and Type of Diversification: Theory and Evidence', 12, 33-48.

Engelsman, E. C. and A. F. J. van Raan (1994) 'A Patent-Based Cartography of Technology', Research Policy, 23, 1-26.

Granberg, A. (1988), Fiber Optics as a Technological Field, Lund: Research Policy Institute, Lund University.

Granstrand, O. (1998) 'Towards a theory of the technology-based firm', Research Policy, 27, 465-489.

Granstrand, O., P. Patel and K. Pavitt (1997) 'Multi-Technology Corporations: Why They have "Distributed" Rather than "Distinctive Core" Competences', California Management Review, 39(4), 8-25.

Henderson, R. M. and I. Cockburn (1994) 'Measuring Competence? Exploring Firm Effects in Phamaceutical Research', Strategic Management Journal, 15, 63-84.

Henderson, R. M. and I. Cockburn (1996) 'Scale, scope, and spillovers: the determinants of research productivity in drug discovery', RAND Journal of Economics, 27(1), 32-59.

Jaffe, A. B. (1986) 'Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value', American Economic Review, 76(3), 984-1001.

Jaffe, A. B. (1989) 'Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers', Research Policy, 18(2), 87-97.

Markides, C. C. and P. J. Williamson (1994) 'Related Diversification, Core Competencies and Corporate Performance', Strategic Management Journal, 15, 149-166.

Metcalfe, J. S. and M. Gibbons (1989) 'Technology, Variety and Organization: A systematic perspective on the competitive process', Research Policy, 4, 153-193.

Miyazaki, K. (1995), Building Competencies in the Firm: Lessons for Japanese and European optoelectronics, Basingstoke: McMillan.

Miyazaki, K. (1999) 'Building Technology Competencies in Japanese Firms', Research Technology Management,(September-October), 39-45.

Montgomery, C. A. and B. Wernerfelt (1988) 'Diversification, Ricardian Rents, and Tobin's q', RAND Journal of Economics, 19(4), 623-632.

Narin, F. (2000), 'Assessing Technological Competencies', in Joe Tidd (ed), From knowledge management to strategic competence. Measuring technological, market and organisational innovation, London: Imperial College Press, pp. 155-196.

Nelson, R.R. and S.G. Winter (1982), An Evolutionary Theory of Economic Change, Cambridge, Mass.: The Belknap Press of Harvard University Press.

Oskarsson, C., 1993, Technology Diversification. The Phenomenon, its causes and effects: Chalmers Unversity of Technology.

Patel, P. and K. Pavitt (1997) 'The technological competencies of the world's largest firms: complex and path-dependent, but not much variety', Research Policy, 26, 141-156.

Patel, P. and K. Pavitt (2000), 'How Technological Competencies Help Define the Core (not the Boundaries) of the firm', in Giovanni Dosi, Richard R. Nelson and Sidney G. Winter (*eds.*), The nature and dynamics of organizational capabilities, Oxford: Oxford University Press, pp. 311-333.

Pavitt, K. (1998) 'Technologies, Products and Organization in the Innovating Firm: What Adam Smith Tells us and Joseph Schumpeter Doesn't', Industrial and Corporate Change, 7(3), 433-452.

Prencipe, A., A. Davies and M. Hobday (2003), The business of systems integration, Oxford: Oxford University Press.

Ramanujam, V. and P. Varadarajan (1989) 'Research on corporate diversification: A synthesis', Strategic Management Journal, 10, 523-551.

Scherer, F. M. (1982) 'Interindustry Technology Flows in the United States', Research Policy, 11, 227-245.

Simon, H. A. (1996), 'The Architecture of Complexity: Hierarchic Systems', The Sciences of the Artificial, Cambridge, Mass: The MIT Press, pp. 183-216.

Valentin, F. and R. L. Jensen (2003) 'Discontinuities and distributed innovation - The case of biotechnology in food processing', Industry and Innovation, 10(3), 275-310.

Valentin, F. and P. Vejrup-Hansen (2004), Udvikling af nye ydelser i rådgivende ingeniørvirksomheder, Copenhagen: Nyt Teknisk Forlag.

Valentin, F., H. Dahlgren and R. L. Jensen (2005), Research strategies in small science-based firms - Effect on performance in Danish and Swedish biotechnology, Paper presented at the Druid Tenth Anniversary Summer Conference, Copenhagen

van Raan, A. F. J. and E. C. Engelsman (1993) 'International Comparison of Technological Activities and Specializations: A Patent-based Monitoring System', Technology Analysis & Strategic Management, 5(2), 113-137.

Verspagen, B. (1997) 'Measuring Inter-sectoral Technology Spillovers: Estimates from the European and US Patent Office Databases', Economic Systems Research, 9(1), 49-67.