EIP

# Bioinformatics or Techbio: Semantic musings from a patent perspective

Bioinformatics has become an increasingly discussed topic in our increasingly digital world. Data analysis has always been part of biology but the increasing need and desire to analyse large data sets and deliver improved automation has caused a discrete discipline to evolve at the intersection between tech and biology. The question remains though, what actually is bioinformatics and would a broader term be more appropriate for most of our discussions in this field?

Merriam-Webster defines bioinformatics as "the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics".

That makes sense, but there are also related terms like computational biology, systems biology, and a new term "techbio". What is the intersection between these terms?

To try and understand this better, we looked at patent publications and patent classification codes in the bioinformatics space to see how the patent offices define this term. The Cooperative Patent Classification system (CPC) was developed by the US Patent Office and the European Patent Office (EPO) and is used internationally alongside the International Patent Classification (IPC).

The CPC defines bioinformatics as "ICT specifically adapted for genetic or protein-related data processing in computational molecular biology" and assigns it the code G16B. Alongside it are Chemoinformatics (G16C) and Healthcare Informatics (G16H).

CPC code ("symbols") can be heavily subdivided and specific. For example, G16B15/10

relates to nucleic acid folding.

p2 Patent publication database Espacenet reports that there are around 22,000 G16B cases, 10,000 G16C cases, and 130,000 G16H cases. However, to be clear, a case is often more than one of these as a patent publication typically has multiple classifications. It should be noted as well that there are many more applications filed for computer devices and methods that are allocated CPC codes A61B, G01N, and G03B. These are typically associated with subject matter such as biomarkers, analysing properties of materials, and imaging. Bioinformatics, at least as far as the CPC is concerned, is a relatively small part of the broader "techbio" space.

We then analysed what classifications were used together on EPO applications published in the last 3 years by using patent analysis tool Patently™, to see what other CPC codes tends to be linked with G16B bioinformatics.

The network graph below shows nodes (dots) each representing a CPC code (at the G16B level of specificity) and the edges (lines) between nodes represents the connection between them. The larger the node, the more patent publications have that code. The thicker the edge, the more patent publications share the two connected code.

If the embed does not work for you, click here https://public.flourish.studio/visualisation/15877811/ to view it in a separate window.

In the graph, you can see G16 as dark blue dots on the top edge of the large mass in the centre. You can click on a node to highlight it and its connections. While selected, there is a pop-up with a link to view information on a node's CPC code on Espacenet. To make analysis of the relevant nodes easier, the other CPC groups in the legend on the top left can also be disabled by clicking them until you find the G16 nodes, and then click the groups again to re-enable them.

From the graph, it is clear that bioinformatics is a relatively small part of the network. Especially compared to behemoths like G06F (Electric digital data processing), A61K (Preparations for medical, dental or toiletry purposes), and Y02E (Reduction of greenhouse gas emissions).

p3

G16B's strongest connections are to G16H (Healthcare Informatics), C12Q (Measuring or testing processes involving enzymes, nucleic acids, or microorganisms) and G01N (Investigating or analysing materials by determining their chemical or physical properties) which is perhaps not surprising given that these classification combinations are currently most likely to have practical commercial applications relating to the analysis of big genomic, proteomic or other data sets.

It becomes clearer what bioinformatics is when we delve into the classification further. CPC further breaks down G16B into 11 subcategories. The most populous of these are G16B 20, 40, and 30. These are:

20 – ICT for functional genomics or proteomics, e.g. genotype-phenotype associations

40 – ICT for biostatistics, bioinformatics-related machine learning or data mining

30 – ICT for sequence analysis involving nucleotides or amino acids

The following graph shows data for all EP G16B bioinformatics families. Cases are commonly assigned to one or more of the major subcategories 20, 30, and 40, as can be seen by the large node sizes and thick connections. Whereas the next largest subcategory: 5 (Modelling or simulations in systems biology), is less linked to the other categories.

If the embed does not work for you, click here https://public.flourish.studio/visualisation/15546083/ to view it in a separate window.

From this, it can be seen that although bioinformatics patents mostly relate to analysing genomic or proteomic data, a few other areas such as simulation, and protein folding are also coming into commercial focus.

EP3426799 (Antibiotic resistance identification) is an example of a G16B patent that is not also in Chemoinformatics (G16C) and Healthcare Informatics (G16H). In summary, this patent claims computer software that takes pathogen genome sequencing data and identifies genes and mutations (e.g. SNPs) within those genes that indicate whether those genomes provide antibiotic resistance or not, and then tells the user its predictions.

Here's claim 1 as granted:

1. A computer-implemented system for identifying antibiotic resistance in pathogens, the system comprising:

-

a gene-resistance module configured to:

p4

- receive as input a plurality of genome sequences, each sequence comprising a plurality of genes,
- generate a gene presence-absence matrix that identifies the genes present in each of the plurality of genome sequences, and
- output a label of resistant or sensitive for each of the plurality of genome samples;
- wherein the gene-resistance module further includes
  - a gene prediction engine configured to identify a set of genes present in a sample of the plurality of genome sequences; and
  - a gene elimination engine configured to remove the identified set of genes from each of the plurality of genome sequences,
  - wherein the gene prediction engine and the gene elimination engine are further configured to iterate the steps of identifying a set of genes present in each of the remaining genome sequences and removing the identified sets of genes from the remaining genome sequences to generate the gene presence-absence matrix;
- a single nucleotide polymorphism-resistance module configured to:
  - receive as input the plurality of genome sequences,
  - identify gene mutations in each of the plurality of genome sequences, and
  - output a label of resistant or sensitive to each identified mutation,
  - wherein the single nucleotide polymorphism-resistance module

p5

further includes:

- an alignment and variant calling pipeline module configured to assemble reads using alignment-based variant calling;
- a variant matrix module configured to generate a variant matrix;
- a single-nucleotide polymorphism resistance association module configured to determine which mutations are responsible for or at least contribute to antibiotic resistance; and
- a single nucleotide polymorphism annotation module configured to annotate identified single nucleotide polymorphisms; and

- an antibiotic resistance module configured to:
  - receive, from the gene resistance module, as input the genes associated with the labels of resistant or sensitive for each of the plurality of genome sequences and, receive, from the single nucleotide polymorphism-resistance module, as input the mutations associated with the labels of resistant or sensitive of the plurality of genome sequences, and
  - identify at least one gene and/or at least one mutation that confers antibiotic resistance and, preferably, the source of a gene that confers antibiotic resistance based on the received labels.

**So what about G16H?**

It would seem that bioinformatics, at least as far as the EPO is concerned, is even more niche than most people mean in common parlance and a look at healthcare informatics (G16H) can be informative for patenting trends in this field as well.

From the first network graph above, you can see G16H (a larger blue dot next to G16B) is also not a large part of the network but it is still bigger and more commonly connected than G16B.

G16H is subdivided into 8 categories, the most populous of which are:

50 – ICT for medical diagnosis, medical simulation or medical data mining; and for

detecting, monitoring, or modelling epidemics or pandemics

p6 40 – ICT for management or administration of healthcare resources/facilities; and management or operation of medical equipment or devices

20 – ICT for therapies or health-improving plans (e.g. for handling prescriptions, for steering therapy, or for monitoring patient compliance).

10 – ICT for handling or processing medical or healthcare data

30 – ICT for handling or processing medical images

This graph shows data for EP G16H health informatics families for the past 5 years. Interestingly, G16H is more generally interlinked among its major subcategories compared to G16B. Albeit 10 (handling medical data) is less commonly linked to 30 (handling medical images), which suggests that G16H patents typically relate to one or the other, not both.

If the embed does not work for you, click here https://public.flourish.studio/visualisation/15589667/ to view it in a separate window.

Looking at published cases in G16H, you can see that they often can be more on the "tech" side than that of G16B. For example:

EP4249045 – Implant communication system and method for communicating with an implantable medical device – Biotronik SE & Co KG – G16H10, 20, 40 and A61N1 (electrotherapy).

EP4250177 – Electronic Health Card – Bundesdruckerei GmbH – G16H10 and G06K19 (record carriers for use with machines)

EP4246320 – Remote management of device user interface content – Welch Allyn, Inc, - G16H40, G06F8 (software management), G06F9 (control units)

However, filtering for cases that are additionally classified for G16B tends to bring us back towards the kind of subject matter we expect for bioinformatics:

WO2023144271 – Method for determining the gut microbiome status – Nestle – G16H20, 50, G16B20, and C12Q1 (biological measuring).

EP4250301 - Method for estimating a variable of interest associated to a given disease as a function of a plurality of different omics data... – Aizoon Srl – G16B20, 25, 40, 45, G16H50, and G06N3 (computing arrangements based on biological models).

EP4184514 - Apparatus and method for diagnosing cancer using liquid biopsy data – Eone Reference Laboratory and GenesisGo Co Ltd – G16B20, 40, G16H10, 50, C12Q1

p7

Looking at the last one EP4184514, in summary, it claims a method that takes sequence data for circulating tumour DNA (ctDNA) in the blood and puts copy number variation information and fragment information into an algorithm to report the stage and origin of cancer. This is a bioinformatics method but more medically focussed. Here's claim 1 of the application:

1. A method for diagnosing cancer using liquid biopsy data performed by a device, the method comprising:

a. acquiring ctDNA sequence information from plasma extracted from blood, and extracting a fragment length and a copy number variations of a chromosome based on the acquired sequence information;
b. extracting fragment reads of P-arm and Q-arm using the acquired sequence information;
c. extracting a copy number variations of mitochondria using the acquired sequence information;
d. inputting at least one of the fragment length and the copy number variations of the chromosome, the fragment read of the P-arm, the fragment read of the Q-arm, and the copy number variations of the mitochondria as an input value of a pre-learned algorithm, and outputting occurrence of cancer as an output value; and
e. inputting at least one of the fragment length and the copy number variations of the chromosome, the fragment read of the P-arm, the fragment read of the Q-arm, and the copy number variations of the mitochondria as an input value of an artificial intelligence algorithm, and outputting a stage and origin of cancer as an output value.

**Conclusion**

The general use of the terms bioinformatics and computational biology interchangeably to mean using computers for biology perhaps needs to become slightly more nuanced when bioinformatics is being considered in patent classification and terminology.

Looking at the patent publication data discussed in this article, we can see that bioinformatics is actually a relatively small area of patent filings, particularly in contrast to how much discussion and interest it is generating in the bio sector as of late. This may be due to the difficulties in patenting pure bioinformatics methods which do not relate to healthcare or other immediate practical applications.

The EPO considers healthcare informatics a separate, and much bigger category which can be more aptly termed Digital Health or Techbio which is the area of most filing activity in this field at the EPO.