Investigating data sources for Biotech firms identification

Stephane Lhuillery[a]

Julio Raffo[b]

Catherine Carpentier[c]

OECD

Workshop on Biotechnology Outputs and Impacts

on 11 December 2006, Paris

[a] EPFL, CEMI

[b] CEPN, Université Paris Nord

[c] INPI, Paris

I. Introduction

The analysis of the impact of biotech on the economy relies on the ability to identify:

- biotech actors and their activity.
- Potential entrants into the biotech research or activities

Statistical surveys are a powerful tool but are bounded by several caveats.

Outline : main sources on Biotech firms

Based on French data we investigate differences between three different sources

- **Identifying biotech firms and biotech public research organizations**
- **Identifying firms likely to be biotech: strategic issues**
- **Helping us to articulate different statistical sources and to improve sampling issues**

Outline

5 questions in this presentation

- How much are we missing with R&D surceys
- How much are we missing with patent statistics
- A bottom-up definition of biotech?
- Identifying the patenting biotech actors
- Identifying potential biotech patentees

Focusing mainly on Independent SMEs

EMPLOYEES GROUP	1 to 19	20 to 499	500 and over	Total
NO	302	124	5	431
YES	29	114	51	194
Total	331	238	56	625

- **Large groups are easy to identify**
- **Their affiliates are not easy to identify as patentees.**

II. How much are we missing with a (census) R&D surveys?

Looking at the likelihood for biotech firms to be respondent to the annual R&D survey, we can conclude that if governments are focused on R&D data, they will get a biased overview:

- Biotech firms without R&D are absent
- Small firms are missing
- Less R&D intense firms are omitted
- Young firms are missed more often
- Biased toward biotech products and processes

Consequences : 2 main aspects

- A statistical aspect: the difference is a problem if people just rely on R&D data where a census is always really hard to realize especially in large countries.
  Any solution to converge?

A feedback loop: The R&D survey must include the biotech firms from the specialized biotech survey if any.

- Policy makers that do not have access to R&D surveys can rely on firms from professional associations taking care with the fact that many small and less high tech firms are missing.

III. How much are we missing with patent statistics?

A. Identifying the patent portfolio of respondent biotech firms

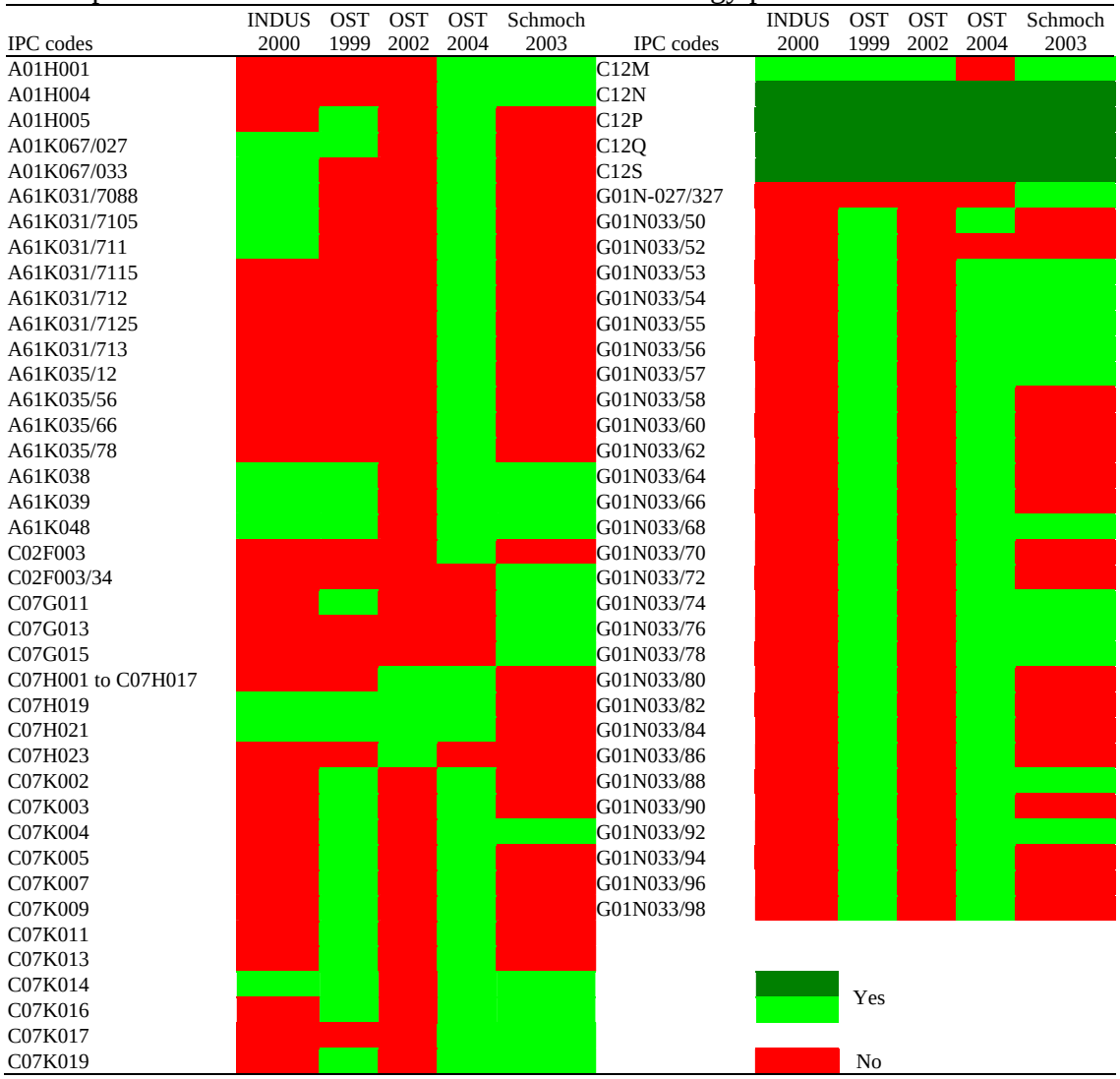
Possible : the Matching firms' name and patentees
Between the pooled survey on biotech matching PATSTAT restricted here to EPAT (1990 – 2006)

Using the N-gram methodology (better than SOUNDEX)

Computing the proximity between each biotech firm and each line in PATSTAT (21 hours on a pentium 5)

Taking care of available different definitions for biotech patents

A comparison between some definitions of Biotechnology patents



B. How many independent biotech firms are identified in EPAT according to the different definitions?

Biotech definitions	Number	Share of firms
OCDE	66	15.9
OST1	65	15.7
OST2	53	12.8
OST3	72	17.4
INDUS1	59	14.2

Among 414 independent French SMEs

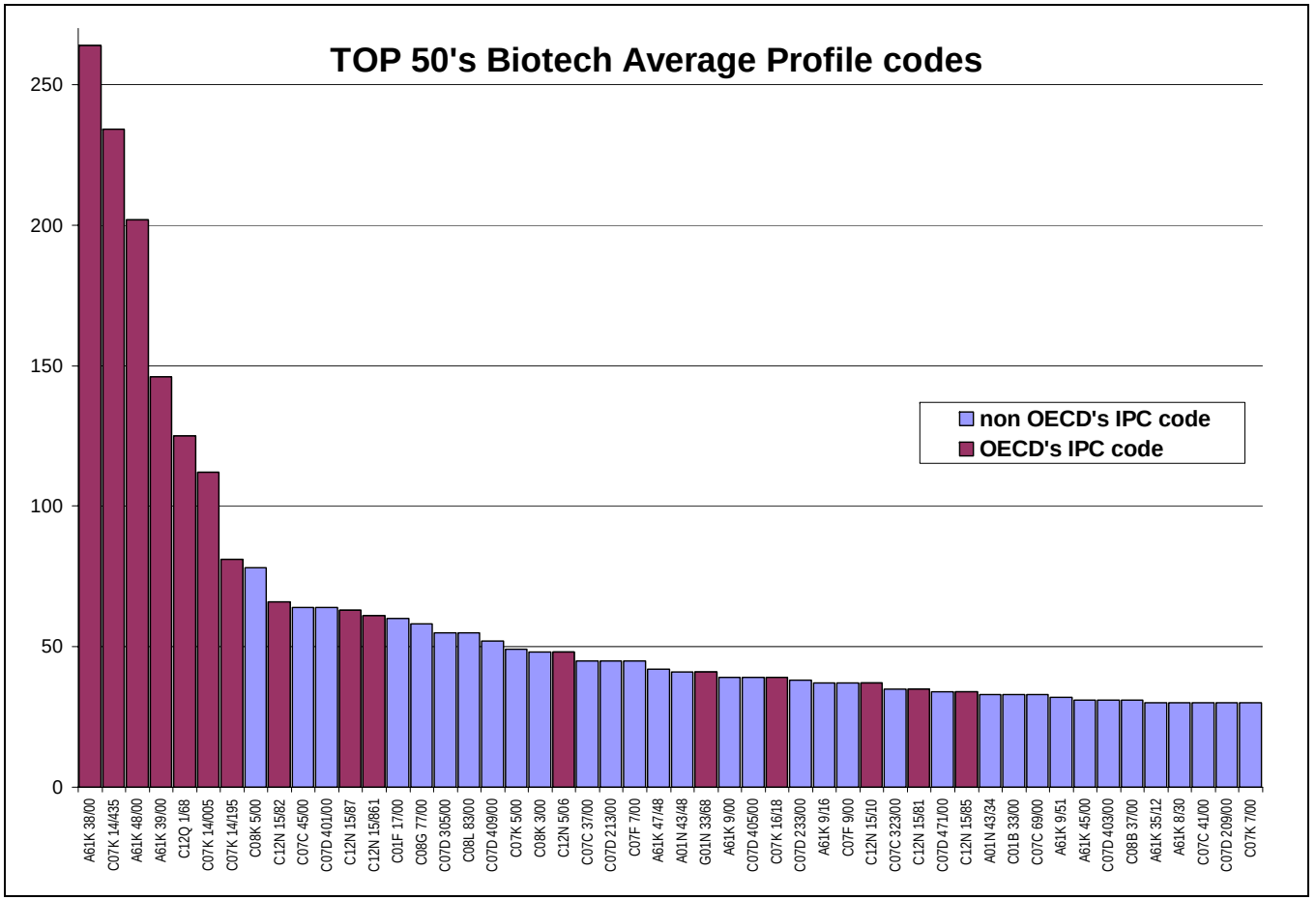
- **Patent data are biased toward:**
  - **Product and process firms**
  - **R&D intense firms**
  - **With R&D services activities**

- **Patent data are not very useful to identify directly biotech firms. The result cast doubts on the patent indicator that is often assumed as reliable in biotech**

- **However, the patent data can be interesting for alternative purposes**

IV. A biotech profile based on patent data

Biotech average IPC codes profile (BAP)



On independent and small firms with biotech process or product

On the differences between bottom-up and top-down definitions:

- **Fields are much broader than the usual expert definitions**

- **How to interpret the difference?**
  - **Experts adopt a special view of biotech activities (tilting the balance toward genetics for example).**
  - **Biotech firms are multitasking and invent in different (complementary) fields**

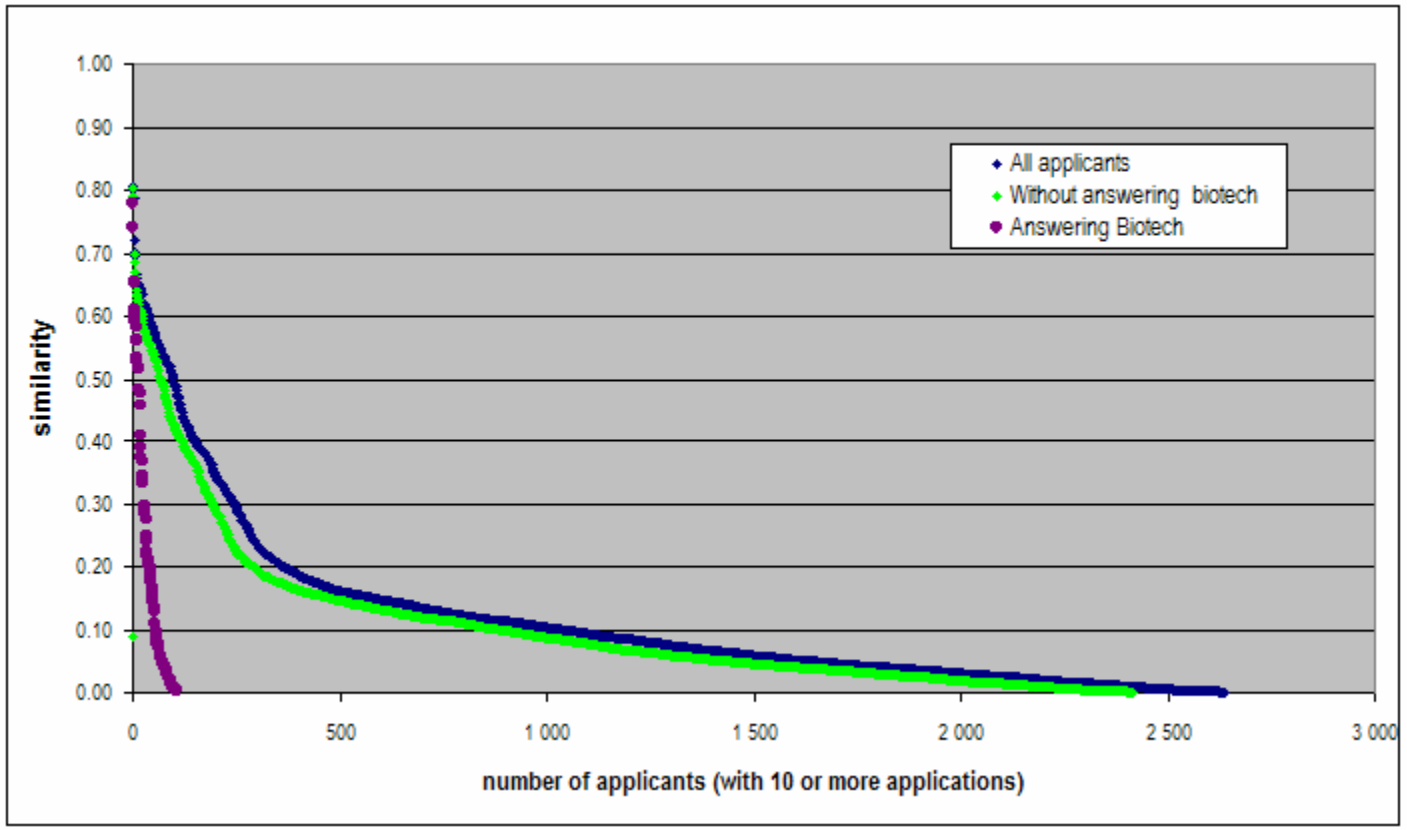
  The response is between the two…

- **Patent data can be used to investigate definitional problems for biotech: a bottom up definition can challenge the expert definitions.**

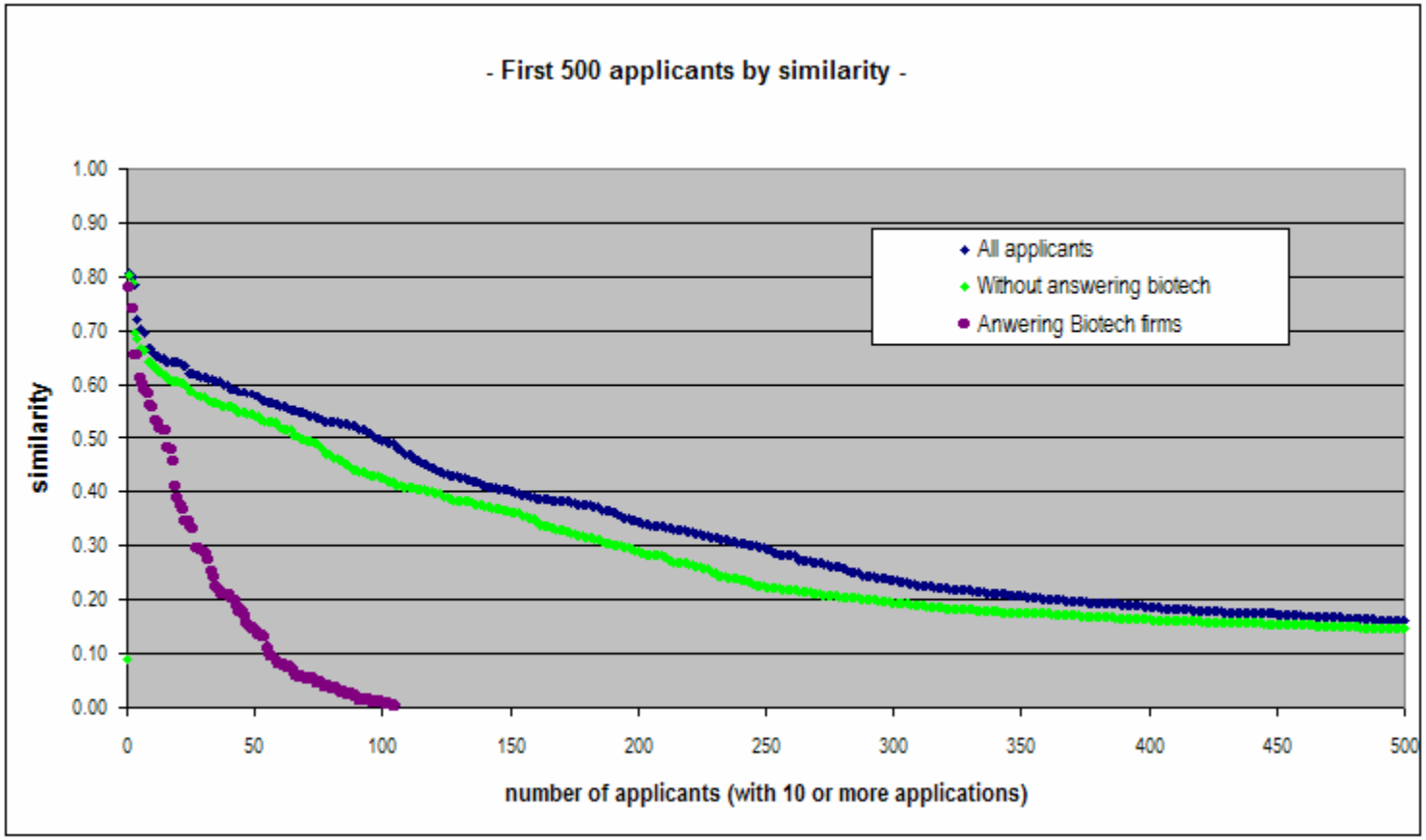
We use the bottom up profile to identify the patenting biotech population.

V. Chasing for biotech actors

A. *A redefinition of the biotech universe*



Including all size, business groups and independent firms
Applicants and inventors are mixed since the applicant field and inventor fields may overlap



Including all size, business groups and independent firms
Applicants and inventors are mixed since the applicant field and inventor fields may overlap

Patent data allow firms to build additional sampling including (The GREEN line):

- Public research organization patenting in biotech (Universities, Pasteur, INSERM, Gustave Roussy)
- Intermediate institutions in charge of biotech (Centre national de la transfusion sanguine, associations and foundations, Genethon)
- biotech firms that are non-respondents to R&D survey and biotech survey
- non-biotech firms that are not far from biotech profiles (potential users or entrants)

A proximity computation based on a bottom-up definition of biotech gives:

- A comprehensive view of patenting actors in biotech or around biotech
- A strategic view of public or private organizations those are likely to enter into biotech research and activities.
- The separation between the two populations is not precise here. The introduction of all firms in R&D or biotech surveys should bring further information on the threshold.

B. Identifying biotech or nearby firms (not answering the biotech survey)

NAMES	Correlation	Count
PROSKELIA	0.673	1
GENCELL SA	0.658	8
CERENIS	0.646	3
CENTELION	0.642	29
SERONO GENETICS INSTITUTE S A	0.641	84
TRANSGENE S A	0.624	43
UROGENE	0.621	2
VAXCONSULTING	0.620	1
INSTITUT MERIEUX	0.606	29
GENE SIGNAL	0.601	4
SB LABORATOIRES PHARMACEUTIQUES	0.590	2
LABORATOIRE LE BRUN	0.587	1
TM INNOVATION	0.587	3
LABORATOIRE EUROPEAN DE BIOTECHNOLOGIE S A	0.570	1
RHONE POULENC	0.538	2881
BIO MERIEUX	0.529	216
INNATE PHARMA SA	0.516	3
ADEREGEM	0.491	2
LABORATOIRE EUROPEEN DE BIOTECHNOLOGIE SA	0.459	2
GENSET	0.452	3
JAVENECH SOCIA TA ANONYME	0.443	2
VETIGEN	0.441	4
CYTHERIS	0.438	2
SANOFI	0.434	1022
MERIAL	0.427	88
ENTOMED	0.423	1
LABORATORIE LAPHAL	0.410	1
PIERRE FABRE	0.402	344
SOCIETE ANONYME ELF SANOFI	0.399	2

NAMES	Correlation	Count
IMMUNO FRANCE SARL	0.401	1
BIO RAD PASTEUR	0.400	16
THERAPTOSIS SA	0.400	3
NEUROTECH SA	0.396	1
L INSTITUT DE RECHERCHE SQUIBB GIE	0.390	2
AVENTIS	0.387	244
IPSEN PHARMA BIOTECH	0.384	1
SCHERING PLOUGH	0.384	14
IDM S A	0.382	1
LAFON PHARMA S A	0.369	1
HOECHST MARION ROUSSEL	0.362	68
VALBIOFRANCE	0.359	7
LABORATOIRES VIRBAC	0.358	6
DIACLONE SA	0.358	2
INNOTHA RAPIE S A	0.358	1
INNATE PHARMA	0.355	7
IMMUNOTECH SA	0.352	9
NOKAD	0.345	1
SOCIETE COTURNIX	0.333	1
GENESIGNAL	0.331	1
PRO SOMA	0.322	1
SOCIETE LEB TECH	0.322	4
RHONE MERIEUX SA	0.321	4
OBE THERAPY BIOTECHNOLOGY	0.317	1
CAYLA	0.313	6
NOVEXEL	0.304	13
GROUPE CELBERT S A	0.301	1
GLAXOSMITHKLINE S A S	0.296	21
TS PHARMA	0.289	2

VI. Conclusion

- THE ASSESSEMENT OF BIOTECH PRIMARLY DEPENDS ON THE ABILITY TO IDENTIFY BIOTECH FIRMS.
- UNCERTAINTY IN THE DEFINITION OF BIOTECH BRING NOISE IN INTERNATIONAL STATISTICS
- IT ALSO MAY NARROW THE VIEW OF BIOTECH FOR POLICY MAKERS

WE ADVOCATE THAT COMPELEMENTARY STATISTIC SOURCES ARE USEFULL TO IDENTIFY BIOTECH ACTORS:

- RD SURVEY, BIOTECH SURVEY AND PATENT DATA CAN BRING INTERESTING VIEWS BUT INCOMPLETE.
- THE THREE SOURCES CAN BE ARTICULATED IN ORDER TO IMPROVE THEIR ACCURACY
- A STRATEGIC VIEW OF FIRMS AND PUBLIC RESEARCH ORGANISATIONS CAN ALSO BE PROPOSED WHEN ACTORS ARE IDENTIFIED AS NOT BEING TOO FAR FROM BIOTECH PROFILES.

Articulating biotech surveys with other data sources (on the firm side only)

Red: Sampling loop
Orange: Definitional loop
Blue : Strategic loop

FURTHER INVESTIGATIONS CAN BE DONE :

- REDUCING THE HETEROGENEITY OF TECHNOLOGICAL PROFILE. A CLUSTER ANALYSIS MAY REDUCE BIOTECH FIRMS TO FEW BIOTECH PROFILES. THE DISTANCE TO THE DIFFERENT CLUSTERS IS A SECOND METHOD TO IDENTIFY BIOTECH ACTORS.

- COMPUTING TECHNOLOGICAL DISTANCE BETWEEN EACH BIOTECH FIRM AND EACH NON ANSWERING FIRMS CAN ALSO BE COMPUTED.

Thank you