Visualization: Final project
Exploring properties of HIV-1 proteins

Alba Refoyo-Martinez, Carlo Baccalaro, Chenhao Wang, Basile Rommes

January 18, 2019

1 Introduction

The project is hosted via github pages under the following url: `https://chenbascaral.github.io/`. The underlying data was accessed via the API of the 'Research Collaboratory for Structural Bioinformatics Protein Data Bank' (RCSB PDB [1] [2], in short: PDB). The PDB is the most comprehensive online archive of protein structures and contains 148037 macromolecules at the time of this writing. It is an important resource to both research and education. For easier data handling, we opted to restrict ourselves to molecules of the Human Immunodeficiency Virus 1 (HIV-1) for the visualization, which amount to a total of 1277 proteins. Besides the 3D structures of the proteins, the database contains several further properties, such as their full name, the protein family they belong to (pfam), the number of amino acids (called residues) that constitute the protein (n_res), their molecular weight (mol_weight), and the date of introduction of the structure into the PDB (date), all of which were used in our visualization.

1.1 Contributions

All members worked together and did a bit of everything. Due to using the Python package Bokeh for visualization, our assignment was very programming intensive. By communicating and meeting up on a regular basis, we worked tightly together on all aspects of the project, including developing ideas for the plots, the design of the visualizations, the progamming and the writing of the report.

1.2 Motivation and Questions

HIV/AIDS is one of the world's major health issues. As it lowers the immune system's ability to protect the body, it is a common underlying cause of death in regions where it is pandemic. Through our visualization, we would like to obtain a better general understanding of the proteins essential to this virus. More specifically:

- Compare the populations of different HIV-1 protein families to each other and see if certain functions require a bigger variety of proteins to perform.
- Determine the relationship, if any, between molecular weight and the number of residues.
- Discover if proteins of the same family are more related by weight or number of residues.
- Determine if any protein family is underrepresented.
- Investigate the growth of the database over the years with respect to HIV-1 protein structures, and gather a better understanding of the current research situation.

The tasks listed above are a perfect fit, because the conclusions can be made directly from a proper visualization without the need for performing any data analysis.

1.3 Target Users

Our visualization targets biologists, HIV researchers and bioinformaticians interested in HIV-1 proteins (the visualization may be expanded in the future to include more species). Such researchers could be interested in a protein's structure, molecular weight, residue length and family. As well, as being interested in the distribution of HIV-1 protein families, such as which families are the most populated and which may be neglected. Our users are expected to have, at minimum, a basic understanding of protein structure and protein properties. A familiarity with the protein database PDB is advantageous as it is strongly linked with our visualization.

2 Data

2.1 Source of Data

The data for this assignment was obtained from the RCSB protein database PDB [2] using the official pypdb API. We first isolated the list of unique IDs for proteins belonging to the species HIV-1, then downloaded their full description, molecular weight, number of residues, protein family and upload date. The data is then stored in a Pandas dataframe in preparation for plotting and manipulation.

2.2 Data Abstraction

Utilizing Munzner's design framework's vocabulary [3][4], our dataset is structured into a static table, with the 1277 items being HIV-1 proteins possessing 7 attributes: PDB id, name of protein, species, protein family, molecular weight, number of amino acid residues, and date of publication to PDB. The first 4 attributes mentioned are of categorical type, and molecular weight, number of residues, and publication date are of type ordered and quantitative, and of sequential ordering.

3 Analysis Tasks

We wish to present and analyze the HIV-1 proteins published in PDB for interested researchers to gain new insights on the proteins and their families.

3.1 Actions and Targets

- Discover distribution of protein families by protein count, so as to ascertain the popularity of protein families and whether some are underrepresented in research and require greater investigation.

- Discover trends, outliers, and correlation in a plot of HIV-1 proteins' molecular weight and number of residues, to learn more about these essential protein properties.

- Present distribution of PDB publication years of HIV-1 proteins, to display the growth of the database over the years with respect to HIV-1 protein structures,and gather a better understanding of the current research situation.

4 Visual Encoding and Interactions

We have decided to visualize the HIV-1 protein data through 3 plots, which are accessible on `https://chenbascaral.github.io/`, and may be switched between via 3 tabs.

4.1 Figure 1: Bar chart of Protein Family Distribution

- Mark: lines.
- Visual variables: length for quantitative value (count), each mark is separated horizontally, ascendingly ordered by counts of the labels (protein family).
- Data: table with 1 category attribute and key corresponding to protein family, and 1 quantitative attribute (count).
- Tasks: compare protein family sizes, look up protein family sizes.
- Scalability: dozens to hundreds of levels for key attribute.

4.2 Figure 1: Table of Protein Family Distribution

- Mark: table rows.
- Visual variables: written values for all categorical and quantitative values, each mark is separated vertically, and ascendingly ordered by counts of the labels (protein family).
- Data: table with 2 category attributes corresponding to protein family ID (key) and name, and 1 quantitative attribute (count).
- Tasks: present protein family IDs, names and sizes.
- Scalability: thousands of levels for key attribute.

4.3 Figure 2: Scatterplot of Molecular Weight vs Number of Residues

- Mark: points
- Visual variables: position (horizontal + vertical), and colouring of points by family.
- Data: table with two categorical attributes, PDB ID (key) and protein family, and two quantitative attributes, molecular weight and number of residues.
- Tasks: finding trends, outliers, distribution, correlation, clusters.
- Scalability: thousands of items.

4.4 Figure 3: Line chart of Published Proteins per Year

- Mark: points connected by a line.
- Visual variables: vertical height for quantitative value, horizontally separated and ordered by key attribute (years)
- Data: table with 1 ordered attribute (key attribute, years) and 1 quantitative attribute(count)

- Tasks: find trend (line connecting the marks emphasizes the ordering of the items along key axis)
- Scalability: hundreds of items.

4.5 Interactions

Specific interaction effects may be turned on and off by clicking on their respective buttons in the top right corner.

- **Navigate** and **wheel zoom in** on items of interest
  - Scatterplot
  - Line chart
- **Select**
  - Bar chart: Able to highlight protein family of interest by selection of bar/row on bar chart or table respectively, unselected families change hue and density.
  - Scatterplot:Click on a particular protein dot to redirect to the protein's PDB webpage.
- **Filter** by box-zoom:
  - Scatterplot
  - Line chart
- **Filter** by legend selection:
  - Scatterplot: Clicking on protein families in the legend to select/unselect them and their corresponding protein dots. The family names in the legend decrease opacity and the dots disappear when unselected.
- **Embed**
  - Bar chart: Hovering over a protein family's bar will present a pop-up of the family's count.
  - Scatterplot: Hovering over a particular protein's dot presents a pop-up containing the protein's 3D structure, PDB ID, family, molecular weight, number of residues, and date of publication.
  - Line chart: Hovering over the points present a pop-up containing the number of proteins published and the year.
- **Details-on-demand**
  - Bar chart: The table on the right displays details of the protein families: family ID, name, and number of proteins. The table row associated with the highlighted family bar turns yellow line.

5 Rationale of Visualization Design

Our visualizations fulfill the Principles of Graphical Excellence and Principles of Graphical Integrity. For instance:

- Principles of Graphical Excellence
  - Complex ideas communicated with clarity, precision and efficiency was achieved by choosing well understood graphs like histogram, scatter plot and line plot that are fitted with interactive elements to deal with scalability issues and confusion that may arise from have a huge amout of data clustered together. See section 4.5 for details.
  - Showing the greatest number of ideas in shortest time with least ink in smallest space comes down to a balance. Since we are studying the proteins, it iss important that all of them are included, so a protein of interest can easily be pinpointed. It is not possible to use less ink to represent the data without compromising this.
  - The data was displayed truthfully as we did not perform any analysis of the data ourselves that could introduce bias, but instead plotted them as they are.
- Principles of Graphical Integrity
  - The principle states that representation of numbers should be directly proportional to the physical measurements. Since the units for number of residues and molecular weight are completely different, not plotting the scatter plot at scale of 1:1 does not violate this.
  - Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity, which is achieved by displaying the full information of each data point when you hover.
  - The graphs are explained with legends and axis labels.

However, our visualization is not without weaknessess and limitations. See section 7 for details.

5.1 Figure 1: Bar Chart

As seen in Figure 1, we have implemented a bar chart to visualize the distribution of protein families by number of proteins, as it is the easiest and simplest method of comparing items based on a single attribute. The bar chart is accompanied by a table of raw data displaying protein family information and a blue underlined hyperlink to the family's entry on the Pfam database [5]. Due to the minimalist nature of our bar chart, there is high data-ink ratio and minimal chartjunk[6]. In addition, the effect of the 'Select' interaction, the yellow highlight of the selected family's row in the table, should not cause an chromatic aberration illusory blur [7][8]. The 'Embedded Hover' interaction generating a pop-up displaying the number of protein family members makes it easier for the user to determine the bar height rather than trying to read the y-axis.

5.2 Figure 2: Scatter Plot

A scatter plot was chosen to visualize the relationship between molecular weight and number of amino acid residues of HIV-1 proteins; the trend line in Figure 2 helps the user determine the two properties' relationship. The dots of the scatter plot, being hyperlinked protein entries, are coloured according to their family, colouration was used to visualize a third protein attribute as colour is a pre-attentive feature, however as there are over a thousand dots in 23 different colours, the preattentiveness of the colouration is less distinct (though the number of colours may be reduced by deselecting families)[8]. Colour was chosen over shape to differentiate protein families as colour is a good distinguishing attribute that is more scalable than shape [8]. Additionally, according to Gestalt Laws of Similarity [9], similarly coloured objects in a plot are perceived to be grouped together. Thus, it is easier for the user to identify clusters of proteins of the same family in the scatter plot if the proteins are coloured by family.

As the scatter plot is quite large with over a thousand data points and distant outliers, there are navigation, zoom, and filter interactions available to facilite the user's investigation of the plot.

5.3 Figure 3: Line Chart

A line chart was used to have a visual representation of the publishing dates for all HIV-1 entries. It also allows us to identify peaks which would be indicative of a large amount of proteins discovered. The 'Embedded Hover' interaction allows the user to see the exact number of protein entries published in that corresponding year.

-

6 Findings

- There is a directly proportional relationship between the molecular weight and the number of residues of HIV-1 proteins as apparent in Figure 4.
- Apparent in the zoomed-in Figure 4 of the scatter plot are clusters of proteins of the same family, as proteins in the same protein family most likely have similar molecular weight and number of amino acid residues.
- The most common HIV-1 protein family is the retroviral aspertine protease (RVP), as shown in Figure 1. Proteases are enzymes breaking down proteins and are involved in a multitude of physiological reactions, a process which the virus is using to destroy infected cells. Since there are a lot of different proteins to target, this family is relatively large. Since proteases are small, they are also easy to crystallize.
- Two proteins (3J3Y and 3J3Q) from the family "Gag_p24" are much bigger in size then the other HIV-1 proteins and appear to be size outliers.
- Since 1989, 1277 HIV-1 proteins had been discovered. About 40 percentage of the protein entries have been identified in the past 5 years (2013-2018).
- Since the foundation of the PDB in 1989, the PDB has been regularly updated. As seen in Figure 3, the year 2013 brought the biggest increase in new entries (118). This boost is due to a breakthrough in crystallography, the most popular technique used to measure the 3D structure of proteins in that year. The new method published by researchers from the University of Tokyo and the University of Jyväskylä was called "porous scaffolding" and made X-ray crystallography of small molecules simpler, faster and more sensitive [10].

The disparity of family size can have one of two explanations:

- Some families encompass a broader range of proteins by nature, and thus are comprised of more proteins.
- Some families are underrepresented in the database due to difficulties in obtaining their molecular structure.

7 Limitations

- The HIV-1 protein families legend's colouring in the scatter plot is not colourblind-sensitive, as we found it difficult to discern between families when coloured using a colourblind-friendly palette, due to having 23 different families and over a thousand dots plotted. The large number of datapoints also results in many overlapping dots when the graph is viewed zoomed out and it is hard to determine which family they may belong to. This limitation is rectifiable by filtering out protein families through the legend or by simply zooming in.

- In the scatter plot, because some colors are similar (e.g. orange and light orange), it may be hard to distinguish between a bulk of overlapping colors of a light color, and a few points of a darker shade of the same color. Again this limitation is partly solved by allowing the user to hide families.

- An improvement on our 'Molecular Weight vs Number of Residues' scatter plot would be to title our legend which was not possible with Bokeh, hence why we described the legend's colouring purpose in the plot title instead.

- While Bokeh is a very versatile framework, having interactions between plots is not possible in a Bokeh standalone document, and instead requires a Bokeh server which we had difficulty implementing into Github pages, our chosen source for publicly hosting the visualization.

- Our entire visualization could easily be extended to the whole protein database, although it poses some challenges in scalability, data management and elegant visualization. Scalability would be a major challenge due to the massive amount of protein entries in the Protein Data Bank (148037 proteins against the 1277 HIV-1 proteins that we have visualized). If extended to the entire PDB database, we would have over a thousand protein families visualized in the bar chart which would be nigh impossible to discern. It would also be much harder to discern overlapping data points in the scatter plot if it possessed over a hundred thousand protein entries and thousands of colours in the family legend.

References

[1] *RCSB PDB.* URL: `https://www.rcsb.org/` (visited on Jan. 17, 2019).

[2] H M Berman et al. "The Protein Data Bank Helen". In: *Nucleic acids research* 28.1 (2000), pp. 235–242. ISSN: 0305-1048. DOI: `10.1093/nar/28.1.235`.

[3] Tamara Munzner. *Visualization Analysis & Design Visualization Analysis & Design.* 2015. ISBN: 9781466508934. DOI: `10.1002/9781119978176`.

[4] Luana Micallef. *Lecture 4 in Visualization - Visualization Analysis & Design.* Nov. 2018.

[5] *Pfam.* URL: `https://pfam.xfam.org/` (visited on Jan. 17, 2019).

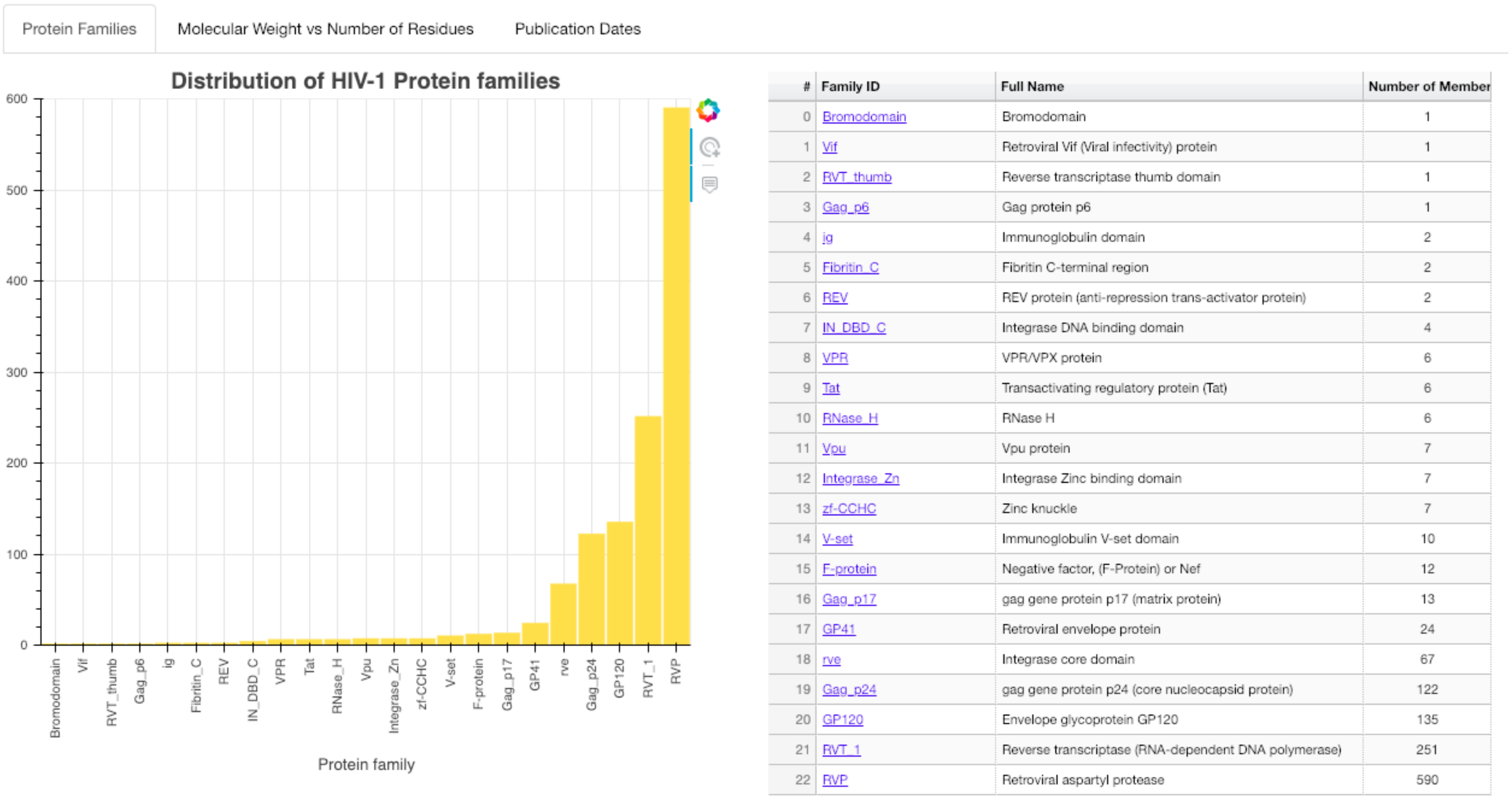
[6] Luana Micallef. *Lecture 5 in Visualization - Visualization Design + Practices.* Dec. 2018.

[7] Luana Micallef. *Lecture 2 in Visualization - Human Visual System 1.* Nov. 2018.

[8] Colin Ware. *Information Visualization: Perception for Design.* 3rd ed. Morgan Kaufmann Series in Interactive Technologies. Amsterdam: Morgan Kaufmann, 2012. ISBN: 978-0-12-381464-7. URL: `http://www.sciencedirect.com/science/book/9780123814647`.

[9] Luana Micallef. *Lecture 3 in Visualization - Human Visual System 2.* Nov. 2018.

[10] Yasuhide Inokuma et al. “X-ray analysis on the nanogram to microgram scale using porous complexes”. In: *Nature* 495.7442 (2013), pp. 461–466. ISSN: 00280836. DOI: 10.1038/nature11990. URL: http://dx.doi.org/10.1038/nature11990.



#	Family ID	Full Name	Number of Member
0	Bromodomain	Bromodomain	1
1	Vif	Retroviral Vif (Viral infectivity) protein	1
2	RVT_thumb	Reverse transcriptase thumb domain	1
3	Gag_p6	Gag protein p6	1
4	ig	Immunoglobulin domain	2
5	Fibritin_C	Fibritin C-terminal region	2
6	REV	REV protein (anti-repression trans-activator protein)	2
7	IN_DBD_C	Integrase DNA binding domain	4
8	VPR	VPR/VPX protein	6
9	Tat	Transactivating regulatory protein (Tat)	6
10	RNase_H	RNase H	6
11	Vpu	Vpu protein	7
12	Integrase_Zn	Integrase Zinc binding domain	7
13	zf-CCHC	Zinc knuckle	7
14	V-set	Immunoglobulin V-set domain	10
15	F-protein	Negative factor, (F-Protein) or Nef	12
16	Gag_p17	gag gene protein p17 (matrix protein)	13
17	GP41	Retroviral envelope protein	24
18	rve	Integrase core domain	67
19	Gag_p24	gag gene protein p24 (core nucleocapsid protein)	122
20	GP120	Envelope glycoprotein GP120	135
21	RVT_1	Reverse transcriptase (RNA-dependent DNA polymerase)	251
22	RVP	Retroviral aspartyl protease	590

Figure 1: Bar Chart and Table of Protein Family Distribution

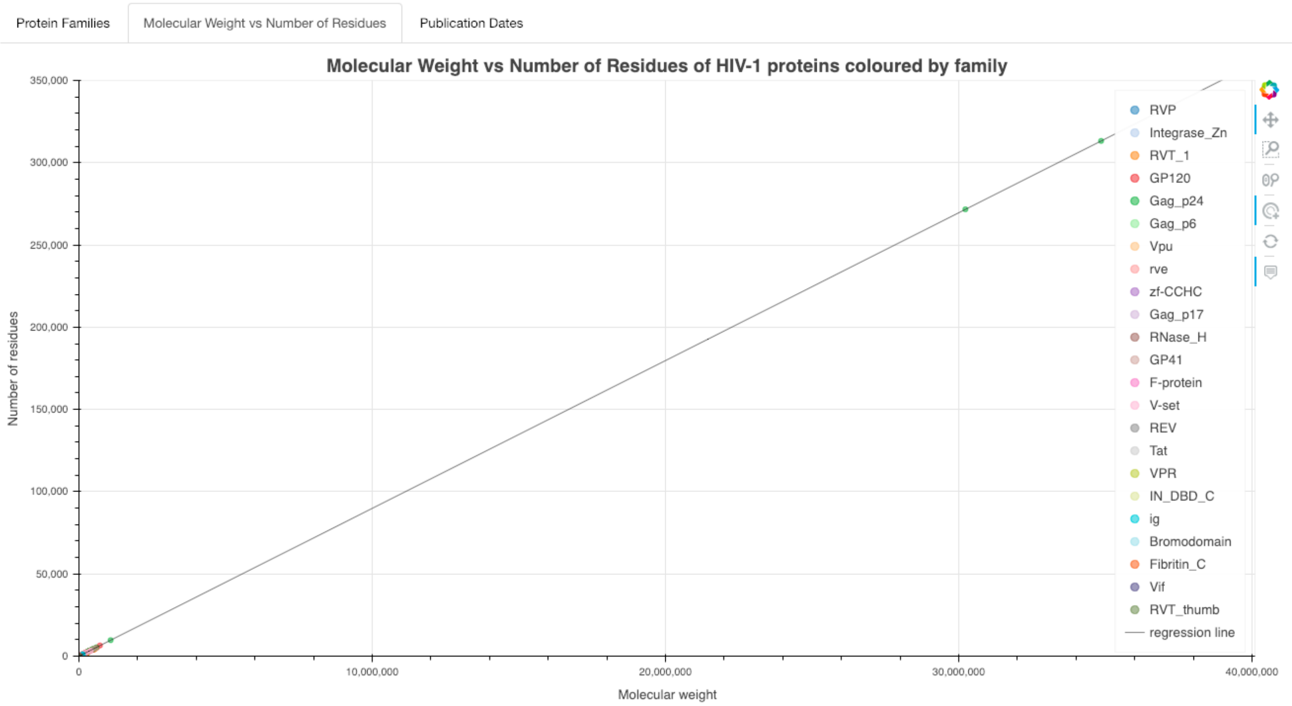


Figure 2: Scatter Plot of Molecular Weight vs Number of Residues

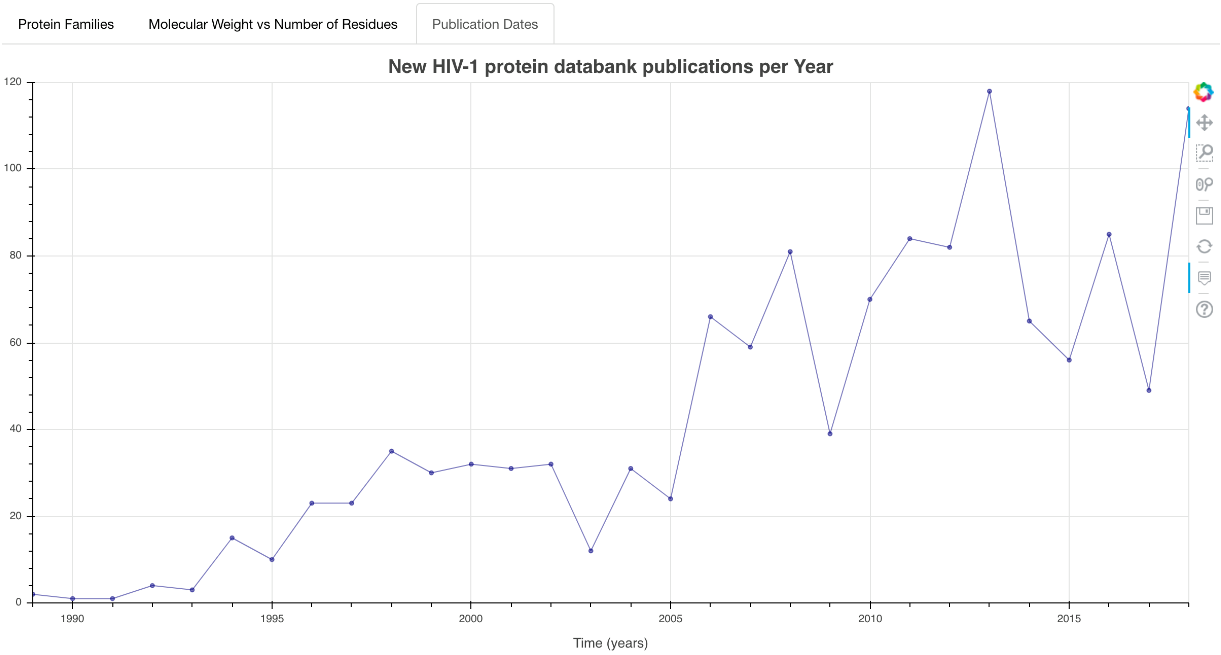
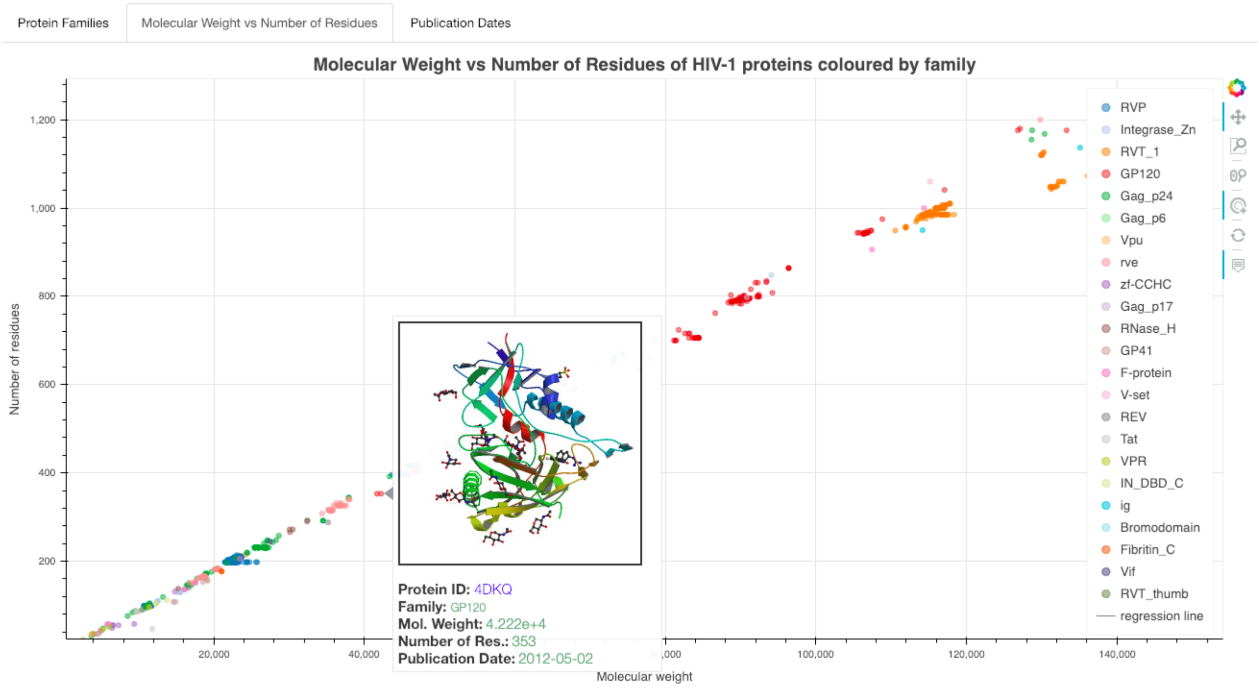


Figure 3: Line Chart of Proteins Published per Year



Protein ID: 4DKQ
Family: GP120
Mol. Weight: 4.222e+4
Number of Res.: 353
Publication Date: 2012-05-02

Figure 4: Zoom in version of the Scatter plot