INTERNATIONAL DESIGN CONFERENCE - DESIGN 2012
Dubrovnik - Croatia, May 21 - 24, 2012.

APPLYING BIOINFORMATICS ANALYSIS PRINCIPLES TO CAD DATA TO BETTER CHARACTERIZE AND IMPROVE THE ENGINEERING DESIGN PROCESS

P. T. Nguyen, M. Steinert, A. Carroll and L. Leifer

Keywords: computer-aided design, design rationale, sequence and cluster analysis, automated data capture, design process

1. Introduction

Computer Aided Drafting/Design, or CAD, is an increasingly important part of consumer and business product design, encompassing ever more design phases. The resulting products can vary tremendously in complexity. Designers use CAD to automate several key processes, such as assembly, version control, verification, and collaboration. There are many use cases for modern CAD software focusing on final design. Modern software suites typically includes prototyping, modeling, proof of concept building, and functional 3D mock up. At the same time, the flexibility of CAD makes it possible to design products from pens to airplanes. During the initial design stages of the Boeing 777, Boeing discovered that collaboration on CAD models was so successful, they scrapped all plans for design verification using physical duplicates [Snyder 1998]. It is due to this increasing flexibility and deployment of CAD systems that we pose the following questions:

1. How can we quantify the processes in which CAD is used in order to gather industry best practices?
2. How can we improve and optimize the ways in which CAD is used?
3. How can we interpret CAD usage data in a way that is coherent given diverse use-cases?

Sung has introduced methods to automate data logging for CAD software, and suggested the high value in knowledge transfer resulting from synthesis of such data [Sung 2010]. We propose to systematically collect CAD data from users and to analyse them algorithmically in order to understand, model and improve engineering design processes. We classify designers in this case to be the engineer users of various CAD software, and are creators of digital artifacts. Our method of analysis is inspired by the field of bioinformatics. Bioinformaticians use analysis programs for a variety of reasons, including calculating alignment trees for protein sequences as part of "phylogenetic tree estimation, structure prediction and critical residue identification" [Edgar 2004]. For example, Edgar's program MUSCLE (MUltiple Sequence Comparison by Log Expectation) uses a 3-stage refinement algorithm to efficiently calculate distances and clustering in order to achieve efficient analysis of large number of sequences. Similarly, we can structure our CAD analysis in the same ways as protein sequences, and thus we need an efficient and accurate approach to analyze thousands of sequences with lengths in the thousands. Such sequences represent individual designers working on individual projects, and the lengths of the sequences represent the chronological series of activities measured. The focus of this paper is to introduce our potential data sources, analysis approach, and our preliminary analysis on a small CAD usage data set.

2. Practical motivation

Our motivation for improving the CAD design process is three-fold. First, CAD is popular and will continue to be an integral part of design in the foreseeable future. If we are able to gather best practices, we would save designers time and money while improving quality. Secondly, design is no longer always co-located, which means multi-team and multi-geography collaboration will create new needs for modern CAD software. Thirdly, we want to create a method of analyzing chronological data that can be generalized beyond CAD in order to treat the entire design process. CAD provides an excellent starting point due to the ease in which we can capture a lot of data. Furthermore, there are several key questions of interest for the various stakeholders around CAD software uses.

1. CAD software developers
    a. How to improve user experience with CAD software
    b. How people are actually using CAD software
    c. How to shape direction for future products in order to stay competitive
2. CAD software users
    a. How to achieve the task at hand easily, effectively, and efficiently
    b. How to use a tool that does not hinder creativity
3. Client companies
    a. How to stay cost effective
    b. How to improve effectiveness of designers without hindering creativity
    c. How to capture best practices

To address the guiding questions above, we propose to thoroughly quantify the CAD-user interaction and to use algorithms to analyze how designers currently use CAD software.

One of the biggest challenges in design research is to capture quantitative data that accurately measure and describe the design process. We define the design processes to be the steps, activities, and physical actions that designers conduct in some chronological order that ultimately lead to the product outcome. With CAD design software, we are able to accurately and seamlessly (automatically) capture the actions and parameters of the tools used. At the same time, the evaluation of outcomes from these processes may be qualitative in nature, whereas the sequences of activities and their correlated timestamp are quantitative. Additionally, it is much more challenging to measure other variables such as the environment, the mental states of the individuals, and the experience. While we recognize that the mental process is among the most important aspect of the design process, we believe that ultimately mental processes will reflect themselves in measurable physical actions. Furthermore, in engineering design, we are particularly focused on activity, rather than creativity.

3. Research questions

This brings us to our research questions. Can we accurately measure the CAD design process in an exhaustive and meaningful manner? We aim to simplify the measurement procedures, or the design problem representation, so that we can minimize the number of variables. We then want to test whether we can generalize this methodology to problems of larger scales.

3.1 Research background

The tie between the design process and outcomes remains very complex. Two designers doing similar things on CAD may be working towards different outcomes. Two designers working towards the same outcome may use two very different approaches. More generally, design has been characterized as a complicated set of interactions within five design stages: Defining the Problem, Needfinding & Benchmarking, Bodystorming, Prototyping, and Testing [Leifer 2010]. Capturing this data implies we are dealing with a chronological sequence of activities. There exists a lot of traditional qualitative approaches to measuring these activities [Malte 2011], [Edelmann 2011]. The coding process include trained personnel watching videos of physical activities and then appropriately grouping the activities and emotions. This allows the researcher to compare respective teams based on their activities and emotions. This qualitative data, along with a measure of team success (based on external evaluation), enable the researcher to answer questions like: What team performance criteria are most conducive to

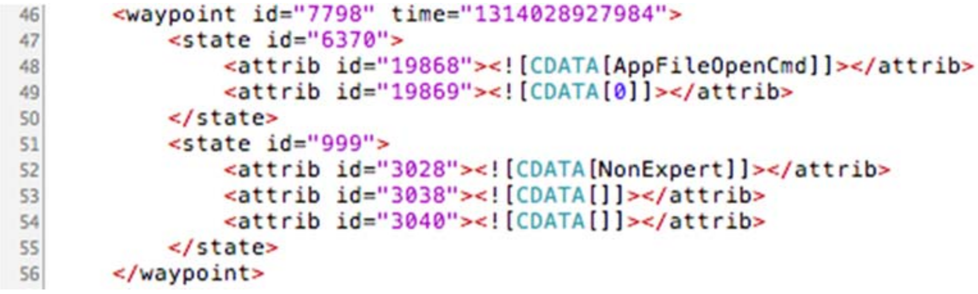
success?
The drawback of some qualitative analysis methods include the subjective coding steps, requirement of extensive training of experts, and a time consuming process (lack of automation). Meanwhile, there are proposed methods to automatically capture CAD data from designers [Sung 2010].

4. CAD activity data

Many CAD companies collect customer-usage data without an explicit plan or strategy for analyzing the data. While surveys asking how designers use CAD tools are important, they typically offer a skewed view of the customers feedback. Additionally, it provides insights into what the customers wish they were doing, or are feeling, rather than what they are actually doing. In contrast to theses existing approaches, by looking at explicit numerical data, we can get a different, and perhaps more relevant insight into what users are actually doing. In particular, we expect to get a view into several key areas.

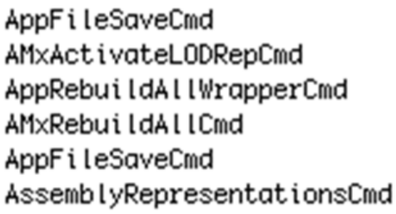
1. How are CAD users actually spending their time
2. What is the activity order in which they spend their time
3. How different users approach similar problems

Though these questions are interesting from a business application perspective, primarily we would be able to answer several key research questions that are of interest at Center of Design Research, Stanford University. In particular, we are interested in how designers works. Traditionally, we monitor them through visual observations, video recordings, surveys, and performance reviews. In a real design process, design insights and thinking can happen anytime and anywhere. In many cases, this is away from the work environment, or during a time that is not possible for us for collect data. For instance, we cannot capture what happens exactly when someone creatively solves a problem while taking a shower or walking in the park [Currano 2011]. Therefore, we are narrowing our definition of design to just the CAD software environment, which is the sole data source. In particular, we collect three types of data: *activity, task complexity, and time.*
The data format is in XML (Extensible Markup Language) which includes action commands, options, and time stamp. External data includes spreadsheets that correlate action commands to activities, and the data sequences to designers' experience.



```
<waypoint id="7798" time="1314028927984">
    <state id="6370">
        <attrib id="19868"><![CDATA[AppFileOpenCmd]]></attrib>
        <attrib id="19869"><![CDATA[0]]></attrib>
    </state>
    <state id="999">
        <attrib id="3028"><![CDATA[NonExpert]]></attrib>
        <attrib id="3038"><![CDATA[]]></attrib>
        <attrib id="3040"><![CDATA[]]></attrib>
    </state>
</waypoint>
```

Figure 1. Example of XML data coming from a user session while using CAD software



```
AppFileSaveCmd
AMxActivateLODRepCmd
AppRebuildAllWrapperCmd
AMxRebuildAllCmd
AppFileSaveCmd
AssemblyRepresentationsCmd
```

Figure 2. After we parse the string of way points with commands, we get a list of commands. Here they are displayed in sequence, without additional contextual information (like time, or options)

The data that we gather are classified as waypoints, each with appropriate commands and time. When we parse this information through extraction of commands using a simple regex search, we are able to get a linear sequence of commands used. While it is of our interest to track and correlate all the state and attributes to a dictionary provided by the CAD company, we start initially by only looking at the command. We will discuss the potential scenarios in which we can gather this data. The current CAD data set that we used for our preliminary tests include around 200 command sequences, tracking a few dozen commands. We supplement the data analysis by including additional non-CAD data sources in order to demonstrate scalability of our analysis methodology.

4.1 Challenges with CAD data

There are three key challenges with using CAD data. First, CAD design quality is a function of the tools as well the proficiency of the operator. Therefore, we have to be careful whether we are measuring the tool, the designer, or the relationships between the designers and the tools. Examples of the latter include training with the tools, previous experience, and other externally enforced constraints. Since CAD design is highly dependent on skill levels, there exists a large amount of variances in experience. Therefore, we would have to collect a large amount of CAD data, as well as relevant contextual information about the designers and their work environment. The second challenge with CAD is dealing with a large number of tools, and the variables associated with those tools. For instance, an extrusion process is described by the magnitude and vectors of the extrusion, the selected surface of the extrusion, and the interactions of the extruded surface with the surrounding components. This information is extremely hard to understand if we look at the information without the proper context. The third challenge is understanding the relationship in the ordering on which the designer interact with CAD. We need to understand the dependencies on the sequence of activities, including the commutative, associative, if any, properties of the sequences. Therefore, our insights are highly dependent on the source of CAD data.

There are several key assumptions we can make.

1. CAD data does not capture all design activities
2. CAD data distinguishes activities at a high resolution with respect to activity and time
3. CAD data offers insights into the outputs of the activities

When comparing multiple design tools, we can use two dimensions: processes & methods, and similarity in outcome. We can constrain the tools and methods used, or constrain the outcomes (Table 1). The following are a few examples of design processes and where CAD fit in.

Table 1. Two dimensions of design constraint

	Doing similar things *(constrained processes & methods)*	Doing different things *(unconstrained processes & methods)*
Achieve similar outcomes *(constrained outputs)*	Machines/robots for manufacturing	Design competitions such as 99Designs
Achieve different outcomes *(unconstrained outputs)*	CAD software like AutoCAD & Inventor	Design consultancies such as IDEO

Loosely, from table 1, we can define that CAD software put a constraint on the activities that a designers may do. For instance, there are a specific sets of tools like drawing, dimensioning, simulation, etc. Through using a linear combination of these activities, over time, the designer is able to achieve a wide variety of outcomes of varying complexity, like a chair or a space shuttle.

As a consequence of the constraint on the set of available tools within the CAD software, CAD data offers a unique opportunity for algorithmic analysis. For instance, we can compare the designer of the chair to the designer of the space shuttle. In essence, they have gone through the same set of tools, but it is the ordering, the sequence, and the number of activities that distinguishes the differences in the outcomes.

4.2 CAD data sources

There are three potential types of data sources for CAD. First, we can collect real usage data from industry professionals who use CAD software for commercial purposes. Second, we can collect data during Quality Assurance (QA) processes, which are internal data generated by the developers of the CAD software. Thirdly, we can create a controlled environment for data, which is to screen and give precise instructions to our targeted CAD users so as to minimize the number of variables we introduce. We will discuss each of these three sources separately, and how we would gather such data.

4.2.1 Data from professionals designing commercial products

The CAD software company we are partnering with has provided a list of industry customers that are willing to allow data collection on their design processes. Whenever a designer uses CAD, the software can automatically collect the sequence of activities, and store it in a well-defined XML file. As the designer proceeds through the creation process, we collect step-by-step data automatically in the background. At the end of each session, the data is exported, and can be used in our analysis. However, we lack contextual information such as what the designers were working on, the experience of the designers, and any measurement of the value of the design work. This dataset is the most diverse in outputs.

4.2.2 Internal data collection from QA processes

The CAD software release cycle includes a series of QA processes designed to test the software for bugs, mistakes, or other rejects. Similarly, the QA process can also simulate how the customers may use the CAD software. QA processes typically include a set of documented procedures or use-cases that a QA engineer would follow. The checklists are built from both experience, product requirements, and other guidelines specific to each company. Some QA processes are simplified so that it tests only touch-points for changes from previous versions. However, in our case, we are collecting data from the simulation of end-users' behaviors, and the QA process includes a comprehensive end-to-end representation of a project. This means that the QA team acts as if it were end-users of the CAD software, which means it has to start the project from outside of CAD by doing things like preliminary design work and product requirements, etc. This type of testing provides us a clear dataset of the *entire* design process, from "start" to "finish". Furthermore, the projects are repeatable, giving us data points that are comparable over time. One disadvantage of this data set is that the simulation is not a perfect representation of end-users. This particular end-to-end QA process typically tests corner cases, and use functions that end-users do not always use. Furthermore, the QA engineers have a more in-depth knowledge about the CAD software than untrained customers. This dataset provides us few, but long, sequences with constrained outputs and inputs.

4.2.3 Experimental setup for controlled data collection

We set up a controlled CAD experiment to reduce the number of variables in the data. First, we set up computer workstations to be used by our participants. In the first part of the experiment, we ask each participant to work towards a particular goal, such as designing a chair. By giving a broad project definition, we permit some creativity in the design process. We collect the sequence of activities that mark the start to finish. We then tie the output to a performance criteria that judges the final outcome of their design. The method to measure performance is flexible, and may come from multiple sources like expert judgement or peer evaluations. Additionally, we can enforce quantitative performance criteria to the design process by taking advantage of included simulation tools in the CAD package. We can measure things such as strain, stress, calculated weight, as well as ratio of weight to carrying capacity. This allows us additional measure of outcomes that can be linked to the design process. Then we modify the rules, and measure everything again. We control for experience by providing training, and calibrate the users. In the second half the experiment, we constrain the sequence of tools used, and leave the product outcome undefined. This allows us to measure the variations in a product outcome despite similar processes. This dataset gives us a large number of short sequences with constrained outputs but less constrained inputs.

5. Using bioinformatics to analyze CAD data

Capturing CAD data automatically can generate huge amounts of data whose analysis is non-trivial. In order to automate analysis and abstract the complexity of design data to useful conclusions, we adapt algorithms employed to analyze DNA sequence data in biology.

There are numerous parallels between biological sequence data and CAD design processes that allow several bioinformatics techniques to map directly. DNA sequence codes for the construction of proteins, which are a linear chain of components which in combination have a specific function. This parallels our interpretation of a design process in which thousands of individual tasks are combined in sequence to generate a final product.

Both gene sequences and design processes evolve over time, with certain core properties held constant because they cannot be changed, unimportant properties which are not under constraint, and areas in-between, which can vary to produce meaningful change, and which represents relevant differences in the function of the gene or design process.

By developing parallels between genes and design processes, we can also make use of the well-developed field of genetic engineering, the intentional modification of a sequence to produce desired results. By correlating the changes in genes or CAD processes to "fitness" – a metric of success or failure of a process – we can generate hypothetical processes that meet their objectives more efficiently [Renner 2003].

In addition, we can apply knowledge of evolutionary biology to understand how innovations in a process are developed and spread over time. Just as we can extrapolate a tree of life from DNA of many species, with design procedures we can track the birth of an idea and follow its adoption and modification in the population of designers.

This also will allow a profiling of how users alter their behavior as they gain experience with CAD. It will be possible to classify the behaviors typically found in new users and observe how their work flow evolves with experience. This may allow the production of tailored program tutorials which more quickly teach a novice how to think like an expert.

In addition, by determining the relatedness of tasks within the program, it may be possible to add a "suggested functions" to menu navigation to more accurately anticipate the next desired actions of the user, both for the purpose of streamlining task production, but also for the purpose of showing a palette of choices at this juncture that he or she may not have considered.

5.1 A brief overview of biological sequence data

DNA encodes information in a linear, base 4 sequence of chemicals which are represented by the letters A, T, G, and C. A degenerate code of triplets of these characters code for 20 different amino acids which are assembled in a linear chain and whose properties determine the chemical function of the resulting protein. These 20 amino acids are represented by most alphabet characters (e.g. both GAA and GAG code for Glutamic Acid or "E"). Of these 20 amino acids, some are similar to each other (both "K" and "R" are positively charged though they are not identical), while others are very dissimilar. Random mutations which occur may be eliminated by natural selection based on whether they occur in a critical region or are a compatible change.

5.1.1 Task clustering – measuring the similarity between processes

The ability to cluster CAD sessions into groups which share common steps is important for subsequent analysis. This allows the distinction between CAD session whose purpose is to create as opposed to modify, or for example, to design a car as opposed to a plane.

We developed a method to cluster the sample CAD data taken from user sessions. This method is based on the use of commands chosen by the user. Each possible command represents a dimension in a high dimensional space. Each user experience is represented as a vector in this space. These vectors are normalized and their distance to each other calculated by cosine similarity.

Clustering is performed via an agglomerative UPGMA method. This method was chosen for its similarity to biological analysis (Top-down approaches such as bisecting k-means tend to perform poorly on highly diverse sequence data). UPGMA clustering is performed by CLUTO [Karypis 2002]. This generates a tree of user sessions in which closer tree branches indicate more similar user experiences (Figure 3). The commands most informative for clustering the sessions are also shown, with high similarity indicated by a darker shade.

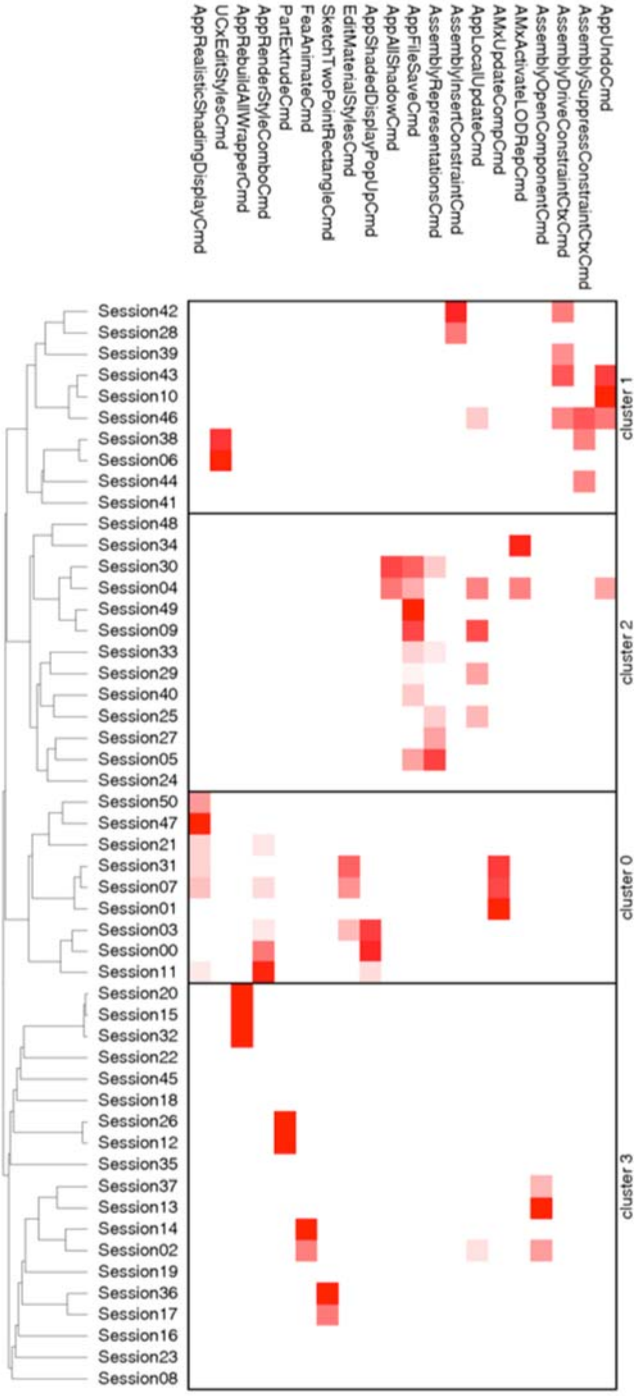


Figure 3. Clustering of CAD user sessions. Numbered sessions are clustered into trees in the vertical axis. Commands informative for clustering are listed in the horizontal axis. Darker shades within a column represent greater use of that command

5.1.2 Comparing processes within tasks

In order to compare two sequences or sets of actions, it helps to know which parts of the sequences are directed at similar purposes. The process of alignment computationally applies heuristics which attempt to match or align similar regions in a set of multiple sequences. This process is driven by a similarity matrix which scores how likely a character in one sequence is likely to have the same function as a character in another. In term of proteins, matching a two "K" gives a score of +5 due to their identity, matching a "K" and "R" gives +2 because of their similar properties, while matching "K" with "I" gives a penalty of -3 because they are dissimilar. The heuristic attempts to maximize the score between a set of sequences by inserting gaps. Gaps themselves have a penalty for opening a gap (-5) and extending one (-1), which prevents an uninformative alignment filled with gaps only resulting in isolated pure matches. The specific heuristics used are complicated and differ across alignment programs and fields (Natural Language Processing has its own process of alignment for example).

We adapted a program called MUSCLE (MUltiple Sequence Comparison by Log Expectation) frequently used to align DNA and protein sequence data to align user experiences on the website Appfluence.com [Edgar 2004]. An alignment of example protein sequence (Figure 4) allows the columns of an alignment to contain a directly comparable region due to the insertion of gaps by the alignment process.

5.1.3 Comparing processes within tasks

We further explain our methodology by expanding our analysis to a larger data set. This is intended to demonstrate that we could analyse when given more data. Due to limitations in availability of CAD data at this time, we chose to work with sequences of pages visited during a browser session from Appfluence.com. By mapping Appfluence.com data to protein characters, we performed a similar alignment of page views using MUSCLE.

There were around 24 accessible pages on the Appfluence.com website, which we translated into amino acid characters (and 4 degenerate characters which are used in biological sequence to represent classes of characters). The browser sessions on Appfluence.com were a chain of visited pages for each user. If a user visited screenshots-videos-about-FAQ-purchase, their session's sequence would read KRSTW, for example. Pages with similar functions were assigned to characters with properties similar to each other, such as assigning the "screenshots" and "videos" pages to characters ("K" and "R"), which both represent positively charged amino acids, and which alignment program's substitution matrix recognizes to value as similar but not identical. This assignment was based on human judgement.

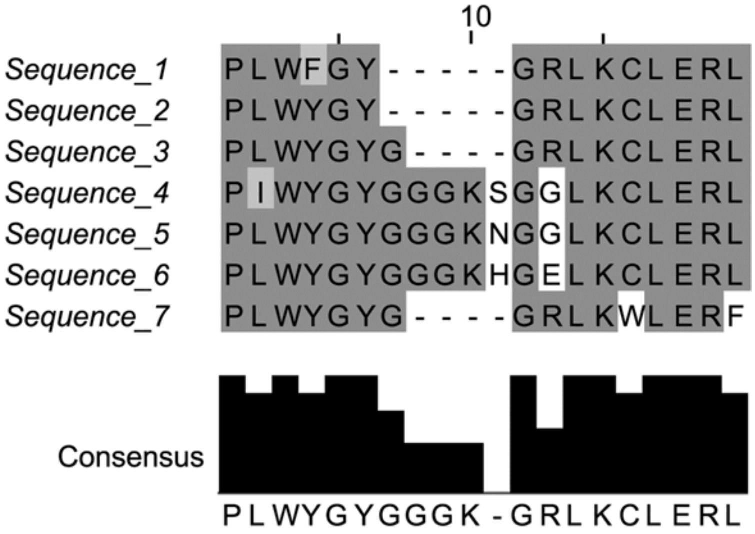

10
Sequence_1 PLWFGY-----GRLKCLERL
Sequence_2 PLWYGY-----GRLKCLERL
Sequence_3 PLWYGYG----GRLKCLERL
Sequence_4 PIWYGYGGGKSGGLKCLERL
Sequence_5 PLWYGYGGGKNGGLKCLERL
Sequence_6 PLWYGYGGGKHGELKCLERL
Sequence_7 PLWYGYG----GRLKWLERF
Consensus
PLWYGYGGGK-GRLKCLERL

Figure 4. Alignment of amino acid sequence from a protein family. Each row is a related protein from a different organism. Letters represent one of the 20 possible amino acid at each position. Dashes indicate computed gaps caused by insertions and deletions over evolution. Characters are shaded more darkly when they are identical or similar to a dominant property in the alignment (e.g. the "I" in Sequence_4 is lightly shaded because "I" is similar to "L")

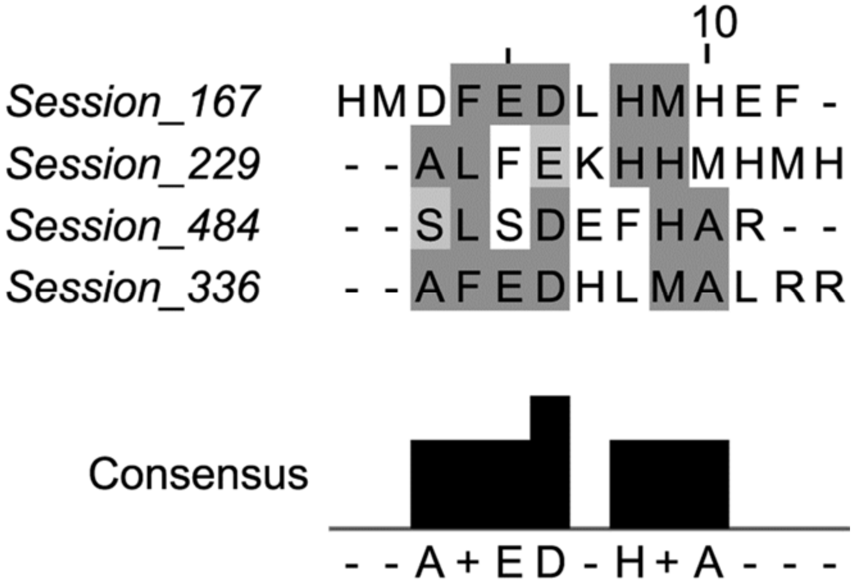

10
Session_167 HMDFEDLHMHEF-
Session_229 --ALFEKHHMHMH
Session_484 --SLSDEFHAR--
Session_336 --AFEDHLMALRR
Consensus
--A+ED-H+A---

Figure 5. Alignment of appfluence browser session. A subset of the alignment of Appfluence.com browser sessions. Pages were encoded as amino acids, attempting to represent pages with similar functions represented as similar amino acids. These were aligned in the same manner as the sequences in Figure 4. These four sessions were taken as a subset of the global alignment by choosing a local region at random. A "+" sign in the consensus track incdicates a consensus property, but without a single dominant character. Shading is as in Figure 4

Although richer information, such as length on a page, was available, in this case we did not make use of this data, treating page hits qualitatively only as a destination on a path through the site. While the browser sessions can be thought of as time series, unfortunately this property does not hold true in biological sequence and so applying the existing alignment tools could not make use of this fact. By extending the method (through building a larger and more complicated substitution matrix) these features of the Appfluence data could be more fully used. Machine learning methods could also use these features in addition to the alignment input to allow more nuanced analysis. It is likely that different design problems will require a different tailored solution, or that a broader framework could be generalized which could automatically assign the proper parameters.

We represented each of 30,000 Appfluence.com user sessions as a sequence and aligned them with MUSCLE. These sessions were the full site data at our disposal. The amount of data required for similar methods will very substantially based on the particular site or problem analysed. This will depends on the homogeneity of the user experience, the number of combinations, and the importance of linear ordering to tasks. Meaningful conclusions about biological data can be drawn in as few as a hundred examples, but it is our estimation that data from human design processes will be more diverse and require far more examples.

The session alignment revealed regions of similarity which represent common experiences between sessions (Figure 4). Colors in both figures represent similar chemical classes (similar pages in Figure 5 were mapped to similar chemical classes), for example both D and E are negatively charged and are colored magenta. Colors in both figures 4 and 5 are based on the clustalX chemical similarity groups: blue, red, green, pink, magenta, orange, cyan, and yellow with no differences in shading. See http://ekhidna.biocenter.helsinki.fi/pfam2/clustal_colours

The sequence alignment also allows for a number of more refined analyses. One such example is improved clustering. The clustering method in Figure 1 used "word" overlap of the constituent characters. The alignment effectively assigns each character in each sequence to a global address, allowing for sequence similarity methods to refine their analysis by directly computing similarity at each position in the alignment. For biological data, this improvement is substantial enough that trees of sequences generally must be constructed through alignment-based methods to pass peer review for publication.

We built trees from the alignment of user sessions on Appfluence.com using the data of the alignment substitution matrix to calculate distance between sequences and applying UPGMA-based clustering. We took a random subset of 10 sessions that ended in exit from the site via a purchase link and 11 sessions which contained a query for support. The alignments of these sequences were used to construct a tree calculating their relation (Figure 6).

This tree indicates differences in the browsing sessions of those who purchase the app are easily separable from those who seek support by this method. This example was a directed experiment aimed to validate the technique. This method could be used on data without preconditions to discover classes of user sessions to characterize the diverse aims of users when accessing a site, or to correlate a desireable outcome (purchasing a product) with a particular user experience in order to promote users to browse in the most favourable way.

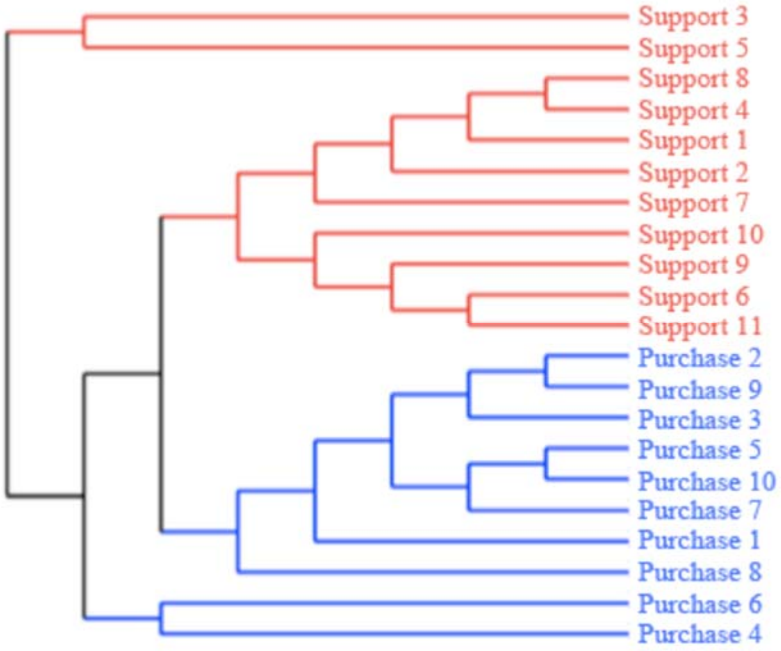


Figure 6. Tree Built from alignment of appluence.com sessions

Distance between Appluence.com sessions was calculated by applying the MUSCLE substitution matrix to a subset of the aligned sessions and UPGMA-based clustering used to construct a tree. There is substantial separation between those sessions which end in a purchase as opposed to those that end in sending a request for support. Analysis of sub-regions of the tree may improve understanding about what users are searching for and how to structure the site in such a way that maximizes positive outcomes

6. Summary and future work

Our systematic approach to gather and analyze CAD data provides tremendous insights into how designers use CAD, the steps in which they take, and their specific behaviors. So far, we have demonstrated that we can understand, and analyze the captured data, and provide activity correlations that can be analysed using open-sourced and proprietary algorithms.

6.1 Industry applications

Even with our preliminary analysis within a limited data set, we are able to see the potential impact of our studies. First, basic clustering gives the CAD tools maker a perspective into commands that happen to have high usage correlation with each other. CAD tools maker can leverage this information by grouping together CAD tools that are correlated, not just by frequency of usage, but perhaps by project success. Second, basic CAD sequence alignment describes different processes that yield similar outcomes. Some of these process pathways may be more efficient over a few desired metrics, such as total time, or number of steps required. It is possible to collect this data as a way to describe industry best practices. How much data we need to capture remains a question we can only answer once we analyse more data. The data set size also depends on the source of the data. CAD data from industry partners working on real projects may be significantly noisier than a lab experiment where we screen users and go through a controlled training process.

6.2 Future work

In future work, we will expand the sequence alignment to CAD data. Since DNA sequence analysis tools typically work with 20 amino acids, we have to either expand the tools to allow for more distincts commands in the sequences, or create a way to group different commands together. Currently, from our XML data, we extracted only the commands that users issued. Then we associated each command to a corresponding letter so that we can use existing tools. By combining additional contextual data, we can associate different sequences with corresponding measure of success, and use that to compare different sequences. Additionally, this method could be used on data without preconditions to discover classes of user sessions to characterize the diverse aims of users through their activities, or to correlate a desirable outcome with a particular user experience in order to promote users behaviors in the most favourable way.

We are currently gathering more CAD data. We are working with one of the global leaders in CAD software in order to instrument industry clients as part of our data collection program. Our strategy is to build a long-term relationship with companies that use CAD extensively. Privacy issues remain a challenge [Sung 2010]. Nevertheless, we are getting promising collaboration opportunities. This enables us to fulfill our initial motivation of measuring designers' effectiveness. We also see an opportunity to set up several controlled experiments where we minimize the number of variables in the CAD design process. This provides us a potential opportunity to compare our approach to methods that use trained experts to code data. Finally, we do not have to stop at CAD data. If we were able to capture all physical (and even mental) activities as a chronological sequence of data points along with parameters and context, we can analyze the data using the same approach we proposed in this paper. We see this as an exciting opportunity to answer questions about best practices in product design as a whole.

Acknowledgement

This research has been graciously funded by the Hasso Plattner Design Thinking Research Program (HPDTRP). It is part of the "AnalyzeD - a virtual design observatory" project.

References

Currano R., Steinert M., Leifer L., "Characterizing reflective practice in design – what about those ideas you get in the shower", ICED'11 (18th International Conference on Engineering Design), 15.-18.08.2011, Copenhagen, DEN, accepted for 2011.

Edelmann J., "Understanding Radical; Breaks: Media and Behavior in Small Teams engaged in Redesign Scenarios", Department of Mechanical Engineering, Stanford University., 2011.

*Edgar, R.C., "MUSCLE: a multiple sequence alignment method with reduced time and space complexity", BMC Bioinformatics, (**5**) 113, 2004.*

Grosskopf A., Steinert M., Edelmann J., Weske M., Leifer L., "Design Thinking implemented in Software Engineering Tools - Proposing and Applying the Design Thinking Transformation Framework", Design Thinking Research Symposium 8 (DTRS8), University of Sydney, Sydney, 19.-20.10.2010, 2010.

Jung M., "Engineering Performance and Emotion: Affective Interaction Dynamics as Indicators of Engineering Design Team Performance", Department of Mechanical Engineering, Stanford University, 2011.

Karypis, G. "CLUTO- A Clustering Toolkit:, Technical Report 02-017. Department of Computer Science, University of Minnesota, 2002.

Leifer L.J., Meinel C., "The Philosophy behind Design", Springer, 2010.

Renner G., Ekárt A., "Genetic algorithms in computer aided design", Computer-Aided Design, Volume 35, Issue 8, July 2003, pp 709-726.

Snyder R.C., Snyder A.C., Sankar C., "Use of Information Technologies in the Process of Building the Boeing 777", Journal of INformation Technology Management, Volume IX, Number 3, 1998, pp 31-42.

Sung R., Ritchie J., Rea H., Corney J., "Automated design knowledge capture and representation in single-user CAD environments", Journal of Engineering Design, 2010.

Martin Steinert, Acting Assistant Professor & Deputy Director, Center for Design Research
Stanford University, Department of Mechanical Engineering, Design Group
Building 560, office 203, 424 Panama Mall, Stanford, CA 94305-2232, USA
Telephone: +1 (650) 862-2996
Email: steinert@stanford.edu
URL: https://www.stanford.edu/group/designx_lab/cgi-bin/mainwiki/index.php/Main_Page