BioTurk: Crowdsourcing the Construction and Augmentation of Biological Pathways

Sasha Koruga

A Thesis
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Master of Science in Engineering

Recommended for Acceptance
by the Department of
Computer Science
Adviser: Professor Olga Troyanskaya

June 2013

За Тихомира и Борку.

Contents

List of Tables

List of Figures

1. Current Pathway Databases

Well-curated biological pathways are highly sought after by researchers, biologists, and bioinformaticians. Several prominent pathway databases have been released into the public sphere in order to quench the demand. For example, KEGG (Kyoto Encyclopedia of Genes and Genomes) provides an assortment of pathways with manually drawn diagrams and a corresponding XML representation for each pathway [9]. SPIKE, another database, provides carefully curated human signaling pathways along with a visualization tool [5]. SPIKE sets out with the goal of making such information amenable to computerized manipulation and analysis by manually gather it and then transforming it into symbolic form by using highly structured languages. REACTOME, yet another manually curated database, attempts to distinguish itself from the pack by making its pathways open-source, open access, and peer-reviewed [8]. The US National Cancer Institute and Nature Publishing Group have also collaborated to create their own Pathway Interaction Database (PID), which consists of curated and peer-reviewed pathways composed of human molecular signaling and regulatory events and key cellular processes [14].

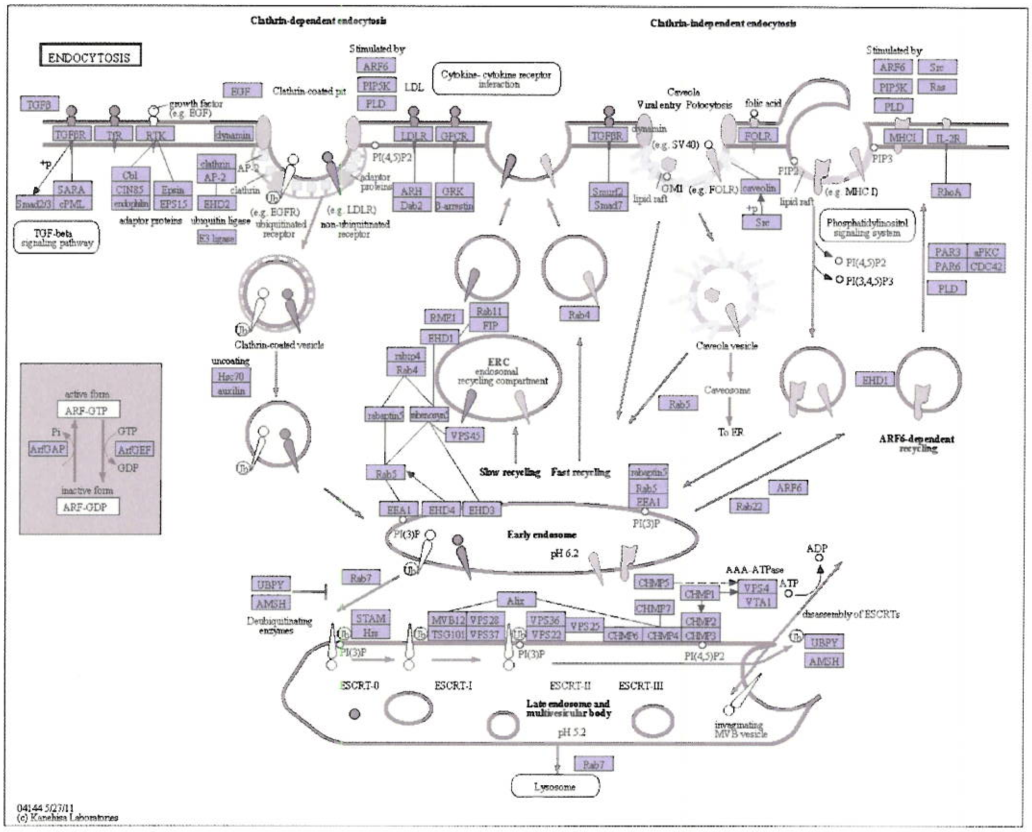
Along with pathway databases, protein–protein interaction databases also end up being highly relevant to our research. Once you know that a particular protein is affiliated with a pathway, finding any other protein that interacts with it can be used to expand the pathway. One such popular interaction repository, BioGRID, was compiled through comprehensive curation efforts [15]. Other manually curated databases include MIPS, the mammalian protein–protein interaction database, which tried to address the concern that most interaction databases were derived from microorganisms rather than mammals [12].

1.1 Endocytic Pathway

KEGG is likely the most widely used pathway database, so we commenced by examining one of their pathways that interested us most at the time. Of genes linked to multiple sclerosis, MHC (major histocompatibility complex) Class II has the strongest association [13]. The endocytic pathway interfaces with the MHCII side of antigen presentation. Proteins phagocytosed by antigen presenting cells are digested and then some of the resulting peptides are loaded onto MHCII; the peptide-MHC complex is then transported to the cells surface. Thus, we took a look at the endocytic XML file and diagram provided by KEGG (Figure 1).

When we examined the KEGG endocytic pathway XML file for protein–protein interactions, we found much to be desired. Out of the thirty-seven connections, one was labeled as "ubiquitination," another one as "phosphorylation," and the remaining thirty-five as simply "binding/association." It quickly became apparent that even these manually curated pathways had much room for improvement.

Another issue that we encountered was that although the KEGG diagrams often represented protein complexes by drawing proteins adjacent to one another, the respective XML files tended to leave those connections unlabeled. While a biologist physically examining the diagram would have not been troubled by this, a bioinformatician who writes software that parses through hundreds of these XML files would have constructed an incomplete picture.



04144 5/23/11
(c) Kanehisa Laboratories

Figure 1: KEGG's endocytic pathway diagram

1.2 RAS Signaling and Tumorigenesis

We carefully examined various pathway databases, and we found that the SPIKE databases tended to have the most complete information when it came to protein-protein interactions. Even though a large chunk of the interactions are still labeled as "physical interaction" or "other," SPIKE tended to have clean and effective XML files. We decided to take advantage of their RAS signaling and tumorigenesis pathway as way to start our users off (Figure 2).

By starting with a known pathway and known connections, we are more likely to capture questions that users will answer in the affirmative. That should make it easier for users as they slowly get acquainted and comfortable with the game mechanics. It also allows us to see if we truly are able to improve upon an already well-documented

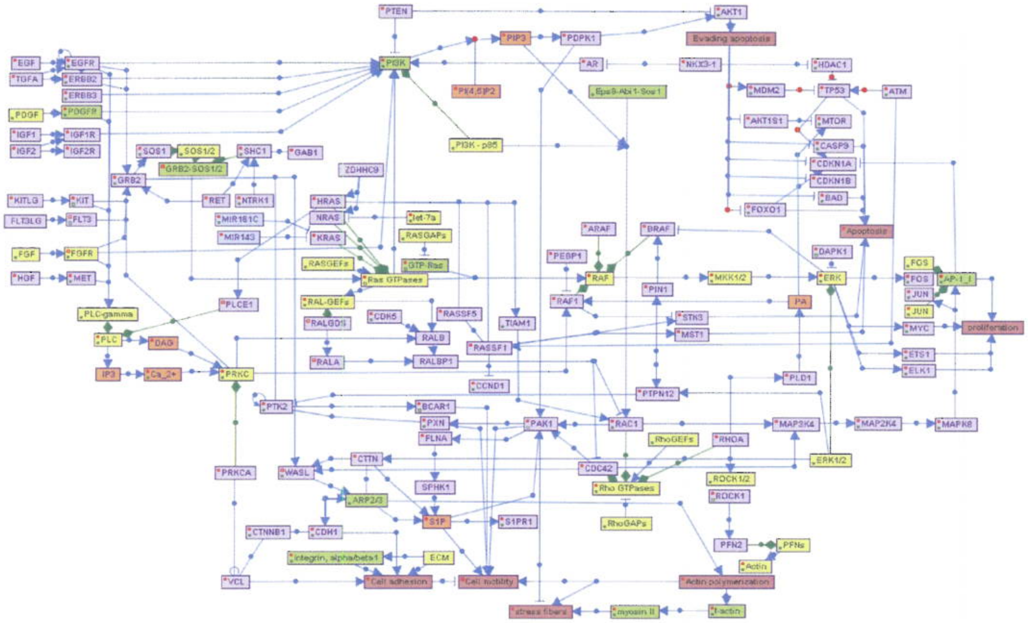


Figure 2: SPIKE's Signaling and Tumorigenesis Pathway

and manually curated pathway. Particularly, we want to see if can find more suitable descriptions for the protein–protein interactions.

2. Sentence Selection

2.1 Data Sources

Our first source of biological literature is pruned from the PubMed abstracts which are freely available online. While we would have liked to have used entire articles as opposed to limiting ourselves to just the abstracts, we are constrained by the fact that our sentences are read across the world by people who might lack the license to view the full article. PubMed currently possess about 22.6 million articles, and approximately half a million are added each year. Its ever-expanding size underscores the need for a crowdsourced solution that would keep up as new literature becomes available. We are at a time where scientific discovery is growing at rapid pace, and lacking the latest data would put scientists at a disadvantage.

Despite PubMed's large dataset, we found that it was just not enough data to find sentences for all of our connections. Interactions often get mentioned in other sections, so limiting ourselves to just abstracts was not going to cut it. Thus, we also added full articles from PubMed Central into our database. PubMed Central is a free full-text repository of biomedical and life sciences journal literature which contains about 2.7 million articles. Conveniently, they make their articles easily downloadable for data mining purposes. Their current data mining set takes up more than 28 GB of computer disk space. Once we added PubMed Central to our dataset, we were finally satisfied with the availability of good sentences.

2.2 Breaking the Data Down to Sentences

We set out with a design goal of restricting what our users see from PubMed down to at most a sentence per question in order to make our game easy and accessible to users without a biological background. Therefore, we wanted to process all our data into simple sentences. However, the task of determining where a sentence begins and ends is not trivial. We start with the obvious: We locate the periods and attempt to use them to carve out the sentences. We quickly notice several exceptions to that rule, such as "et al.," "etc.," and "Fig. #." We run into quite a few of these, so we just started building a database of exceptions to handle them. We also look for some signs of sentence structure that should always be there: a sentence that starts with a capital and a capital letter after the period. These quick rules and small database of exceptions seem to be able to handle almost all of our data. It also runs very quickly, which is important given the sheer size of the data that we are processing.

Since we are looking for sentences that contain protein-protein interactions, we know that there must be at least two proteins in the sentence and at least one interaction term. We use UniProt to create a list of protein names to search for [2]. When you take into account all the protein names and aliases, we ended up with a list of 402,939 names. When we conducted our search, we found that we had too many hits with just a few protein names that overlapped with regular words. Our list included: AND, FOR, WAS, THIS, NOT, CAN, FIG, LARGE, MANY, PER, AREA, SPATIAL, CASE, LINE, YEAR, CARE, IMPACT, HOW, KEY, LOG, LTD, COPY, MED, and SHE. We threw out those protein abbreviations in order to avoid many poorly chosen sentences. Throwing out 24 names out of 402,939 was an insignificant amount of proteins, but made a large impact on the size of our dataset.

For the interaction term, we use the interaction terminology constructed by Park et al., which consists of 124 terms related to protein-protein interactions. We cut-off a select few of the letters at the end of each interaction term in order to make them function as stems (e.g. for "phosphorylation," we store "phosphorylat," so that it catches "phosphorylation," "phosphorylates," and "phosphorylated"). With this interaction term as a requirement in addition to the two proteins, this greatly reduces the amount of sentences in our dataset.

2.3 Ranking Sentences

The goal of our ranking algorithm is to find sentences which are both easy to answer and likely to be answered in the affirmative. If users can understand a sentence and are confident enough to agree that an interaction is occurring, then they are generally inclined to enjoy the game more. Difficult sentences, on the other hand, frustrate users, and users who are not forming interactions due to being serve poor sentences start to believe that they are doing something wrong. Since we cannot fulfill our goal if users do not play our game, we have decided to place a great deal of effort on designing a ranking algorithm that fetches the right sentences.

Our algorithm for ranking sentences is now the focal point of the project and thus under heavy development. Currently, we take into account several factors to determine a sentence's rank, including:

- **The length of a sentence** – We found that the shorter a sentence, the more likely it is that it is going to be easy to read. Furthermore, shorter sentences by their nature do not have a lot going on. Therefore, any short sentence with two protein names and an interaction term is likely going to have the interaction term reference the two proteins. The longer the sentence, the more likely that the interaction term is referring to something other than the two proteins in the sentence.

- **The number of capital case characters** – While the PubMed abstracts tend to be clean text data, the PubMed Central articles often have metadata which slips through our filters. In order to avoid showing those weird sentences, we penalize sentences based on the amout of capitals characters.

- **The number of non-alphanumeric characters** – For reasons similar to the above, we penalize sentences that have too many non-alphanumeric characters. Certainly we expect a few punctuations, but anything excessive will result in huge downward shifts in the sentence's score.

- **Proximity of interaction term and the two proteins** – Even if a sentence has a bit too many characters, as long as the interaction term placement in the sentence is close to the two protein names, it is very likely that that sentence will be answered in the affirmative.

- **The interaction term is a verb** – If the interaction term is a verb (e.g. phosphorylates as opposed to phosphorylation), the sentence tends to be both easier to read and more likely to be answered in the affirmative.

- **The interaction term is in-between the two proteins** – When the interaction term is positioned between the two proteins, it is more likely to follow the predictable "<subject> <verb> <target>" sentence structure.

- **The number of protein names in a sentence** – When there are a lot of protein names in a sentence, it becomes more likely that the interaction is referring to at least one of the non-target proteins. We thus penalize sentences with too many protein names.

- **The article section that the sentence originated from** – Some sections, such as the abstract, are more likely to contain easy-to-read and easy-to-parse sentences.

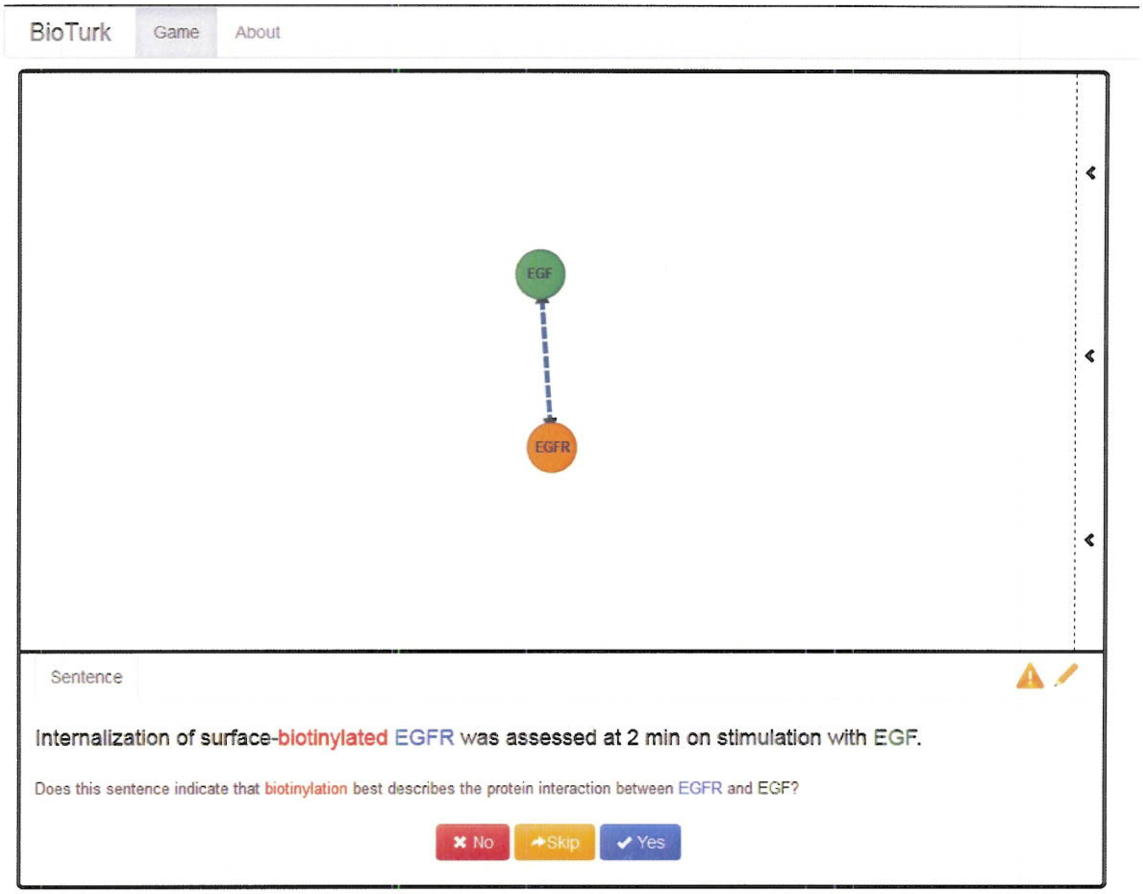
3. User Interface

3.1 Basic Interface

The basic interface provides a user with a sentence mined from either a PubMed abstract or from a full article in PubMed Central (Figure 3). All sentences provided to the user will have two protein names (one colored with blue font and the other colored with green font) and a protein interaction term (colored with red font). When given the aforementioned sentence, a user is asked if the protein–protein interaction term that is highlighted corresponds to the two highlighted proteins. Our hope is that users without biological knowledge will be able to simply use their knowledge of English grammar to answer these questions to the best of their abilities. They are presented with only three options per sentence: yes, no, and skip.

3.2 Discreet Options for Regular Users

During testing, we have found that users with biological expertise sometimes want to change the interaction term or the selected proteins before answering the question in the affirmative. For regular users, however, giving them extra options could cause them to feel overwhelmed. Even with our simplified interface and limited options, regular users tend to find the questions very challenging. Giving these users an extra array of options and tools would likely result in even more confusion. Our primary



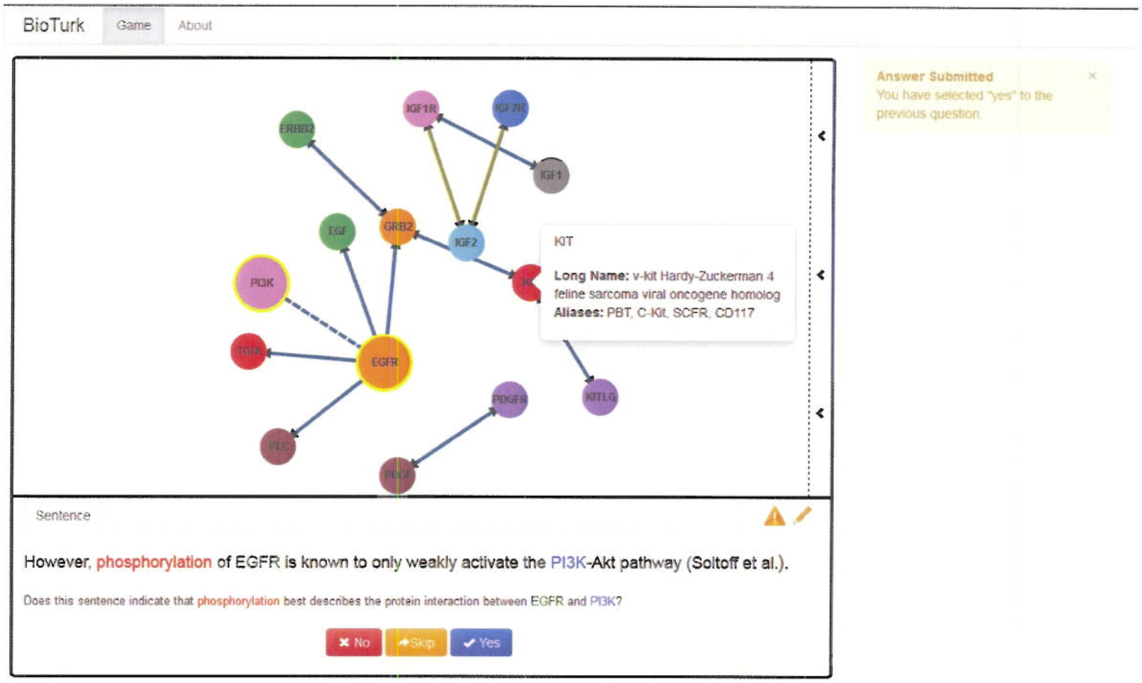
Internalization of surface-biotinylated EGFR was assessed at 2 min on stimulation with EGF.

Does this sentence indicate that biotinylation best describes the protein interaction between EGFR and EGF?

Figure 3: The basic interface

objective is to make sure that users stay with us and continue to enjoy the game while potentially learning about biology; we do not want to scare them away.

There are a few special options and informational tidbits in the interface for regular users. These features are purposefully made discreet in order to adhere to our design standards for these users. Users can gain more information about a protein by hovering their mouse over the protein circles in the network diagram (Figure 4) or the protein names in the question (Figure 5). Performing this action will provide the user with extra information concerning that protein, such as the protein's long name and the aliases that that protein might possess. For example, hovering the mouse over "EGFR" will make text appear at the pointer's location, which explains



Answer Submitted ×
You have selected "yes" to the previous question.

KIT

Long Name: v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
Aliases: PBT, C-Kit, SCFR, CD117

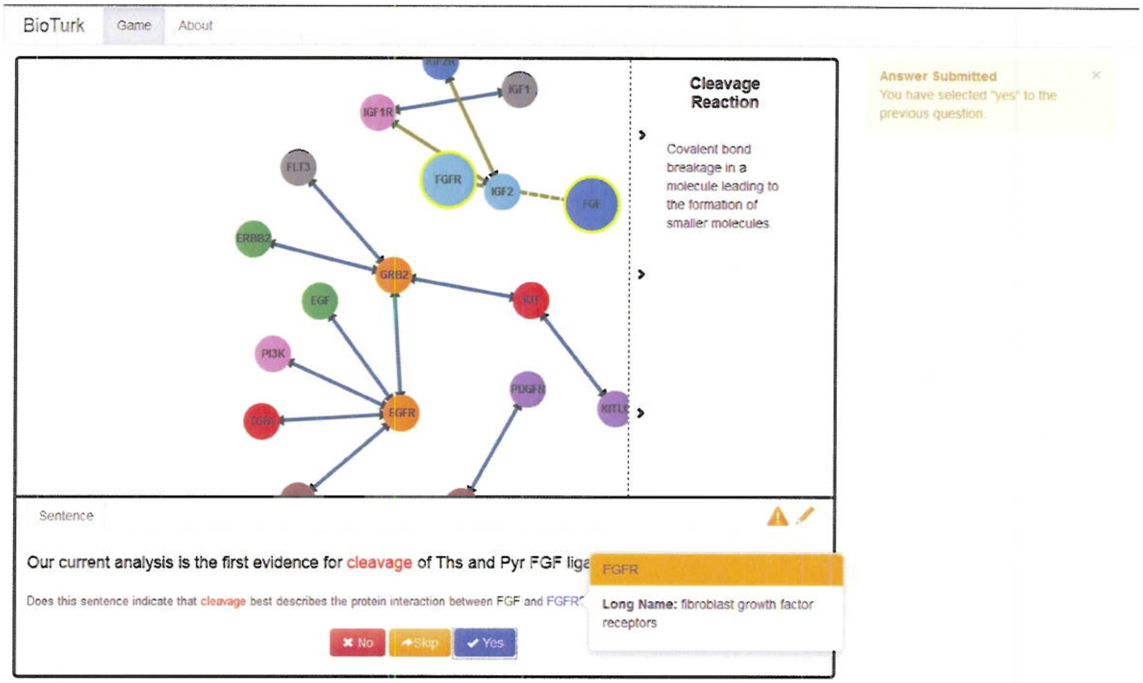
However, phosphorylation of EGFR is known to only weakly activate the PI3K-Akt pathway (Soltoff et al.).

Does this sentence indicate that phosphorylation best describes the protein interaction between EGFR and PI3K?

Figure 4: A user hovers her mouse over the KIT protein in the network map

that the long name for this protein is, "epidermal growth factor receptor," and that the aliases for this protein include ERBB, ERBB1, HER1, PIG61, and mENA. Alias information is actually quite useful, as some sentences will use an alias. We always change the protein names in the question to reflect the protein's alias in the sentence, but we leave the network diagram unchanged. We instead always use the most common name for the protein in the diagram. We decided that changing the name in the diagram with each sentence would be too awkward. This does lead to an occasional disconnect, where users see a protein name in the sentence that does not correspond to the diagram. However, our testing has shown that this does not lead to confusion, as users tend to focus more on the sentence and question at hand rather than the diagram.

Another slightly hidden feature is the panel to the right, where users initially just see three tiny arrows pointing to the left (Figure 6). Users can click on these arrows to expand a panel which holds information about the protein–protein interaction



Answer Submitted

You have selected "yes" to the previous question.

Cleavage Reaction

Covalent bond breakage in a molecule leading to the formation of smaller molecules.

Our current analysis is the first evidence for cleavage of Ths and Pyr FGF liga

Does this sentence indicate that cleavage best describes the protein interaction between FGF and FGFR?

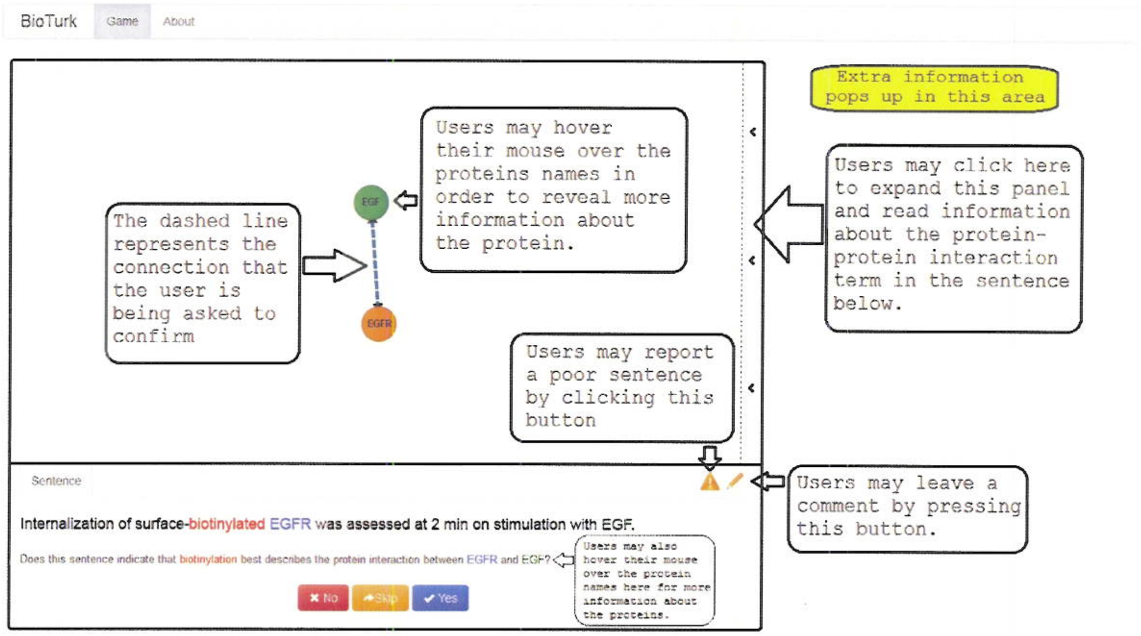
FGFR

Long Name: fibroblast growth factor receptors

Figure 5: A user hovers her mouse over the FGFR protein name in the question

term that is currently highlighted in the sentence below. This panel automatically changes with the sentence to reflect new interactions. We have previously had this panel always remain on display, but we have since decided that this was too much information for a new user. Instead, we hope that users will discover this option after they feel more comfortable with the game. The information about the interactions is meant to help educate the users so that they may be better equipped to answer the questions, and also so that they may walk away from this game with a better understanding and appreciation of biology. If we could pique a user's interest in biology now, they may eventually come back to us as expert users.

Since our ranking algorithm will occasionally present the users with awkward or nonsensical sentences, we allow users to report sentences. By clicking on the little hazard symbol, the user will automatically receive a replacement sentence for the current protein–protein interaction. The reported sentence will be stored in our database so that we can review it and adjust our algorithm accordingly.



Extra information pops up in this area

Users may hover their mouse over the proteins names in order to reveal more information about the protein.

Users may click here to expand this panel and read information about the protein-protein interaction term in the sentence below.

The dashed line represents the connection that the user is being asked to confirm

Users may report a poor sentence by clicking this button

Users may leave a comment by pressing this button.

Internalization of surface-biotinylated EGFR was assessed at 2 min on stimulation with EGF.

Does this sentence indicate that biotinylation best describes the protein interaction between EGFR and EGF?

Users may also hover their mouse over the protein names here for more information about the proteins.

Figure 6: Basic interface with annotations

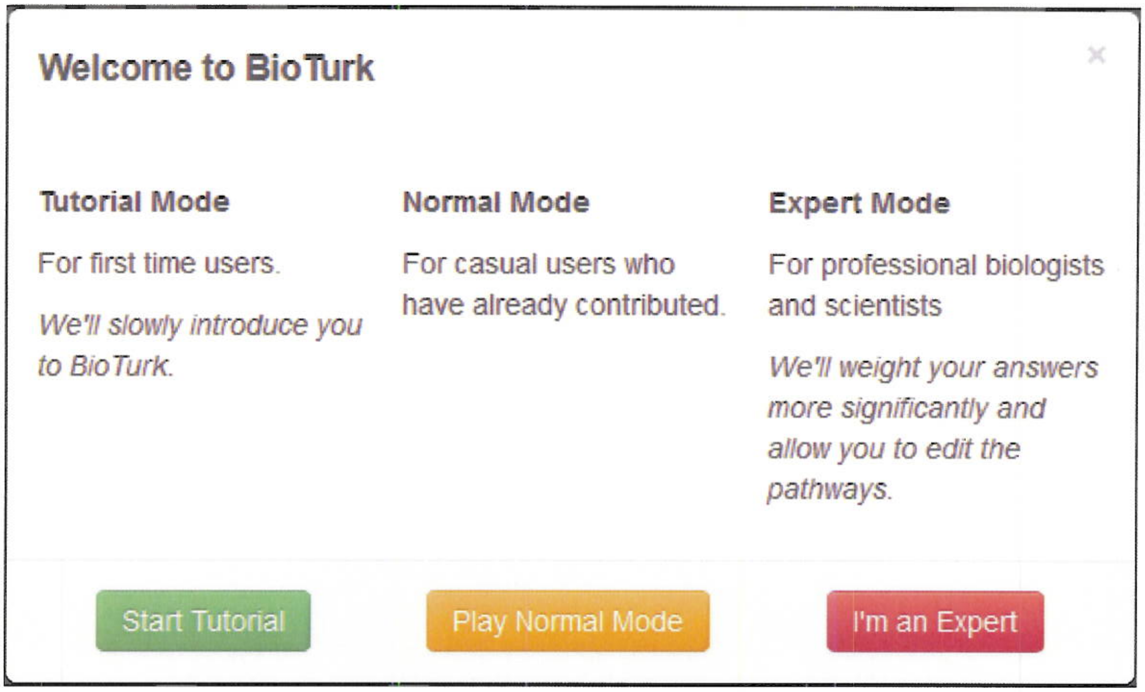
Occasionally users have some basic biological knowledge, but not enough to be considered experts. Those users may choose to leave a note for us on any particular sentence. By clicking on the pencil icon, users will be given the ability to write a comment to associate with that sentence. That comment will be stored in our database for later review. For example, we had users report to us that a sentence with "autophosphorylation," means that a protein phosphorylates itself. Those users would then select "no" for the question, because they knew that phosphorylation did not occur between the two proteins in the sentence. In response to that feedback, we have adjusted our algorithm so that sentences with "autophosphorylation" do not show up again.

3.3 Previous Games & Motivation for Expert Mode

The ESP Game was created by a lab at Carnegie Mellon University [17]. In this game, two users are presented with the same picture and then asked to guess what the other user has written. This game has been immensely popular, and the researches have been able to use it to label a massive amount of images. The fact that they had a thriving player base has allowed them to achieve their goal, and it shows the importance of making sure a game is fun and accessible. This task is much more arduous in biology.

A lab at the Scripps Research Institute has created four gene games with the general goal of collecting data [16]. They do not have separate interfaces for regular and expert users; instead, they suggest that regular users scour the web for information about genes and diseases in order to play their games. We tried all four of their games, but needless to say, the amount of web research required to answer even one question, let alone finish an entire game, makes all four games essentially inaccessible to anyone but field experts. Solely relying on experts, who tend to be few in numbers and often constrained by demanding schedules, is unlikely to yield the necessary amount of data to generate reliable predications. Unless a person creating this type of game is willing to settle with a low amount of data, it is necessary to design a game that would both attract and retain regular users.

For regular users, the sentences from PubMed already tend to be intimidating. This is supported by feedback from our MTurk tester, where they almost unanimously stated that they found the questions difficult. Even students from Bioinformatics and Computational Biology have reported to us that the sentences are often relatively difficult to answer. We have been working intensity on coming up with a methodology to rank sentences by ease and complexity. In other words, we try to present users with sentences that are easy to read and which they are likely to answer in the affirmative. Since this primary task is already arduous, we do not want to add more complexity to the interface for these users who will undoubtedly make up the majority of our player base.



Welcome to BioTurk

Tutorial Mode

For first time users.

We'll slowly introduce you to BioTurk.

Normal Mode

For casual users who have already contributed.

Expert Mode

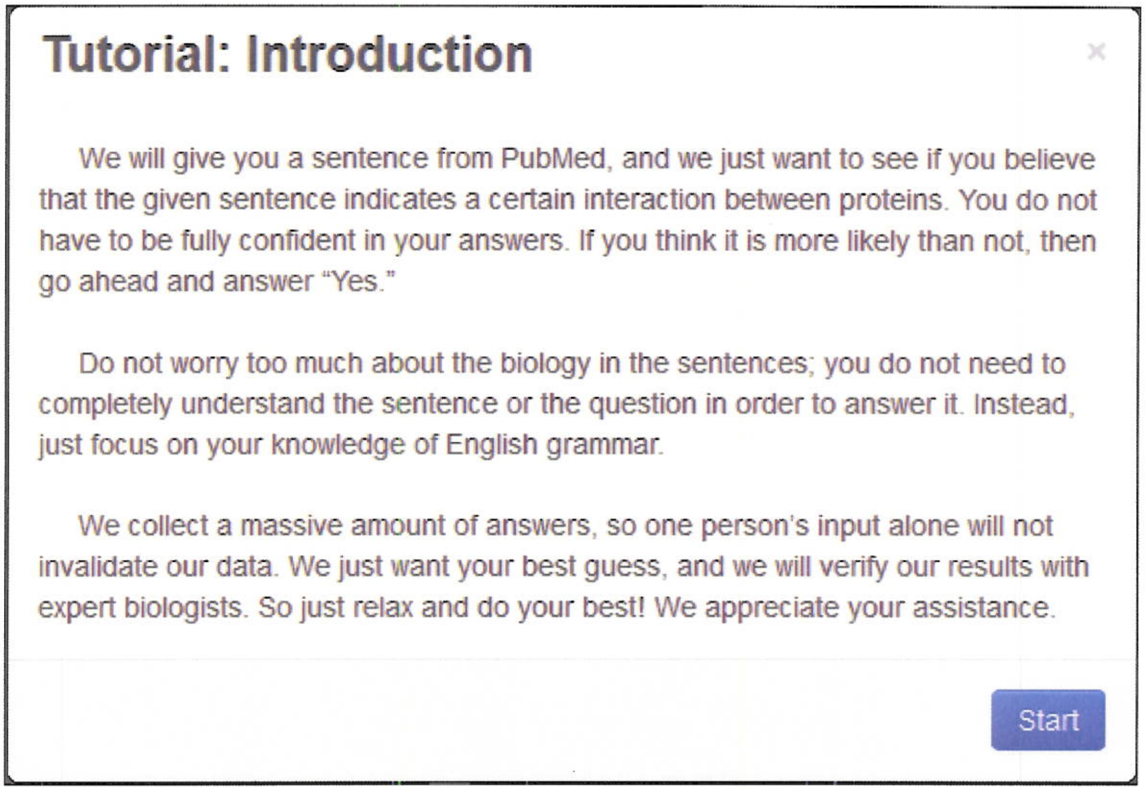
For professional biologists and scientists

We'll weight your answers more significantly and allow you to edit the pathways.

Figure 7: This prompt first appears when a user launches the web application

At the same time, to not leverage the knowledge of expert biologists by granting them greater control would be myopic. Thus, it becomes apparent that by providing two interfaces, one for regular users and one for expert users, we get the best of both worlds. Expert biologists have the freedom to use their knowledge to help us rebuild our network of proteins; regular users, on the other hand, help us answer the easy questions, identify contested questions that should be presented to biologists, and affirm new protein–protein connections purposed by biologists.

We, of course, do not expect that all of our users will correctly answer every single sentence. Instead, our hope is that if a good amount of users answer a question in the same way, we can be rather certain that they are correct. For sentences where we are unable to get regular users to agree, we can present those sentences to expert biologists and see if they arrive to a consensus. Currently, for testing and display purposes, anyone can play as an expert. Our plan, however, is to limit expert mode to verified users. Since we are able to trust these users, we do not need to query many



Tutorial: Introduction

We will give you a sentence from PubMed, and we just want to see if you believe that the given sentence indicates a certain interaction between proteins. You do not have to be fully confident in your answers. If you think it is more likely than not, then go ahead and answer "Yes."

Do not worry too much about the biology in the sentences; you do not need to completely understand the sentence or the question in order to answer it. Instead, just focus on your knowledge of English grammar.

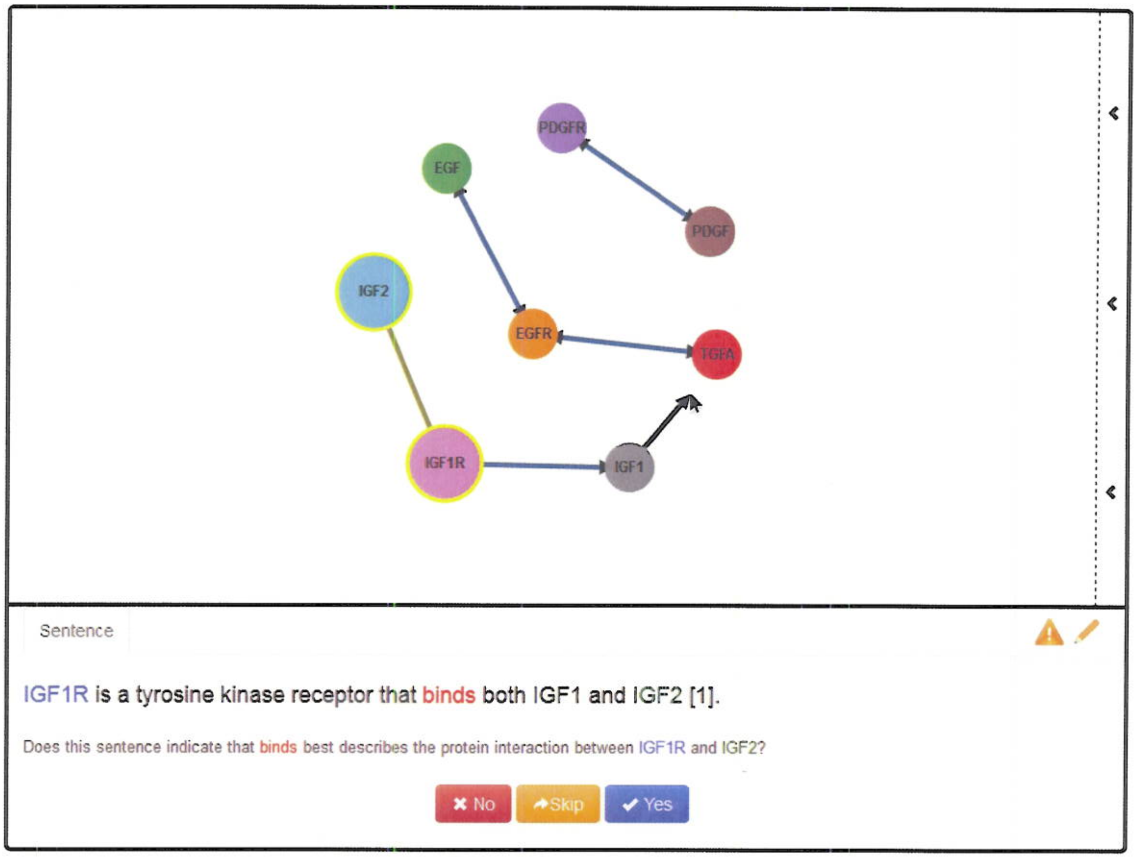
We collect a massive amount of answers, so one person's input alone will not invalidate our data. We just want your best guess, and we will verify our results with expert biologists. So just relax and do your best! We appreciate your assistance.

Figure 8: Explanation for Tutorial Mode users

of them with the same question to form a consensus. Instead, even just two biologists giving the same answer would be conclusive for us.

3.4 Expert Mode Interface

When users first arrive to our web application, they are greeted with a friendly prompt that explains and allows them to choose between our three user modes: Tutorial, Normal, and Expert (Figure 7). After that, they are prompted with the completely optional step of associating a name and/or email address with their answers. Following that, users who have selected Tutorial Mode will get an additional explanation (Figure 8). We found that regular users were concerned that they were invalidating our data or that they did not have the biological knowledge to answer any of our questions. The extra explanation that they receive in Tutorial Mode attempts to ease



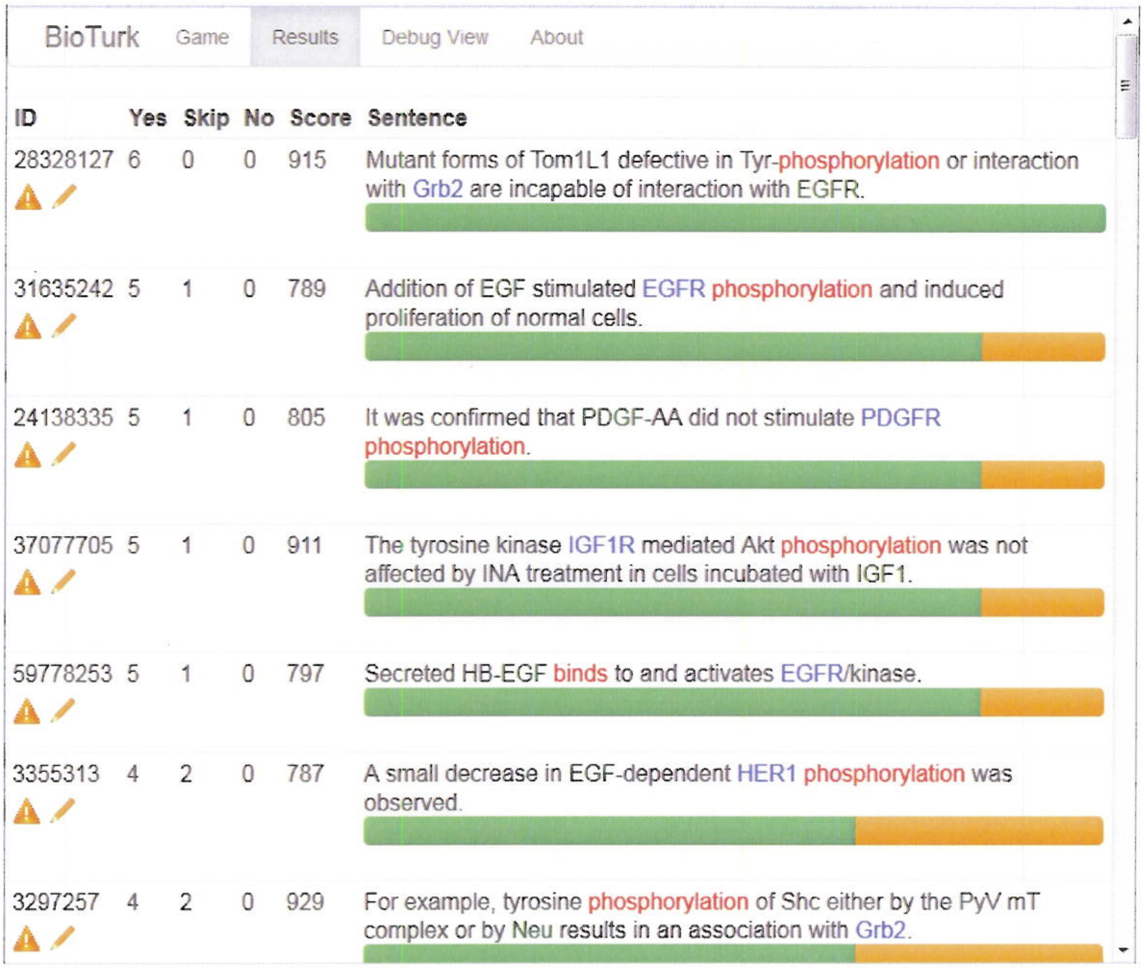
IGF1R is a tyrosine kinase receptor that binds both IGF1 and IGF2 [1].

Does this sentence indicate that binds best describes the protein interaction between IGF1R and IGF2?

Figure 9: The expert user has clicked on IGF1 is dragging the mouse over to TGFA to connect the two proteins

these concerns. We also start these users off with a set of hand-selected sentences that are relatively straightforward to answer in order to slowly ease them into the game.

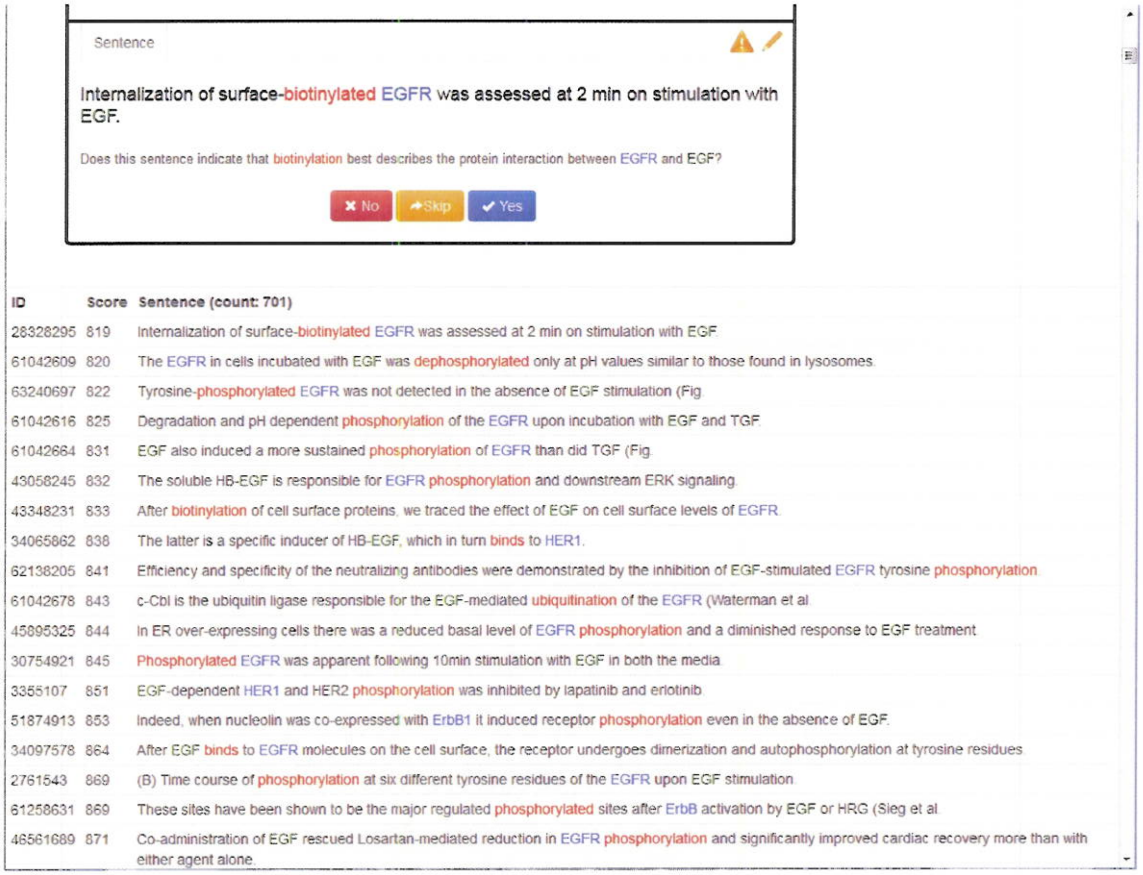
While regular users can only assert or rebuff the connection that the current question is considering, expert users can use their mouse to click and drag between two proteins in the network map in order to create a connection (Figure 9). Once a link is established, the user is prompted to select the interaction occurring between those two proteins. Additionally, if an expert user wishes to create a connection with a protein that is not currently displayed in the network map, she can simply click on any empty space in that map to create and name a new protein.



ID	Yes	Skip	No	Score	Sentence
28328127 ⚠ ✏	6	0	0	915	Mutant forms of Tom1L1 defective in Tyr-phosphorylation or interaction with Grb2 are incapable of interaction with EGFR.
31635242 ⚠ ✏	5	1	0	789	Addition of EGF stimulated EGFR phosphorylation and induced proliferation of normal cells.
24138335 ⚠ ✏	5	1	0	805	It was confirmed that PDGF-AA did not stimulate PDGFR phosphorylation.
37077705 ⚠ ✏	5	1	0	911	The tyrosine kinase IGF1R mediated Akt phosphorylation was not affected by INA treatment in cells incubated with IGF1.
59778253 ⚠ ✏	5	1	0	797	Secreted HB-EGF binds to and activates EGFR/kinase.
3355313 ⚠ ✏	4	2	0	787	A small decrease in EGF-dependent HER1 phosphorylation was observed.
3297257 ⚠ ✏	4	2	0	929	For example, tyrosine phosphorylation of Shc either by the PyV mT complex or by Neu results in an association with Grb2.

Figure 10: The results page allows expert users to see how the regular users are voting on sentences

Expert Mode users are also given greater access to our data. For example, experts can go to the "Results" page to see a list of all the sentences that regular users have rated (Figure 10). The list is sorted by our confidence that a sentence is reliable, and a colored bar represents the proportion of users voting "yes," "skip," and "no" in green, yellow, and red respectively. Expert users can also directly leave comments or report sentences from that list. Lastly, expert users are provided with a debug view, where they can see all the sentences that we are considering for the currently selected protein–protein connection (Figure 11). These sentences are sorted by our scoring function that attempts to determine the easiest sentence that will yield an affirmative response. With all these extra features, we hope that the experts will feel both welcomed and unconstrained.



Internalization of surface-biotinylated EGFR was assessed at 2 min on stimulation with EGF.

Does this sentence indicate that biotinylation best describes the protein interaction between EGFR and EGF?

ID	Score	Sentence (count: 701)
28328295	819	Internalization of surface-biotinylated EGFR was assessed at 2 min on stimulation with EGF.
61042609	820	The EGFR in cells incubated with EGF was dephosphorylated only at pH values similar to those found in lysosomes.
63240697	822	Tyrosine-phosphorylated EGFR was not detected in the absence of EGF stimulation (Fig.
61042616	825	Degradation and pH dependent phosphorylation of the EGFR upon incubation with EGF and TGF.
61042664	831	EGF also induced a more sustained phosphorylation of EGFR than did TGF (Fig.
43058245	832	The soluble HB-EGF is responsible for EGFR phosphorylation and downstream ERK signaling.
43348231	833	After biotinylation of cell surface proteins, we traced the effect of EGF on cell surface levels of EGFR.
34065862	838	The latter is a specific inducer of HB-EGF, which in turn binds to HER1.
62138205	841	Efficiency and specificity of the neutralizing antibodies were demonstrated by the inhibition of EGF-stimulated EGFR tyrosine phosphorylation.
61042678	843	c-Cbl is the ubiquitin ligase responsible for the EGF-mediated ubiquitination of the EGFR (Waterman et al.
45895325	844	In ER over-expressing cells there was a reduced basal level of EGFR phosphorylation and a diminished response to EGF treatment.
30754921	845	Phosphorylated EGFR was apparent following 10min stimulation with EGF in both the media.
3355107	851	EGF-dependent HER1 and HER2 phosphorylation was inhibited by lapatinib and erlotinib.
51874913	853	Indeed, when nucleolin was co-expressed with ErbB1 it induced receptor phosphorylation even in the absence of EGF.
34097578	864	After EGF binds to EGFR molecules on the cell surface, the receptor undergoes dimerization and autophosphorylation at tyrosine residues.
2761543	869	(B) Time course of phosphorylation at six different tyrosine residues of the EGFR upon EGF stimulation.
61258631	869	These sites have been shown to be the major regulated phosphorylated sites after ErbB activation by EGF or HRG (Sieg et al.
46561689	871	Co-administration of EGF rescued Losartan-mediated reduction in EGFR phosphorylation and significantly improved cardiac recovery more than with either agent alone.

Figure 11: The debug view allows expert users to see all the sentences that we have considered for the current interaction

4. Evaluation

After we conducted our May 8th survey, we wanted to see how well the users answered the questions posed to them. We picked two sentences at random to look at. This is the first sentence we studied:

> TGFA is closely related to epidermal growth factor (EGF) and binds to the EGF receptors (EGFR) as a ligand.

Users were asked:

> Does this sentence indicate that binds best describes the protein interaction between TGFA and EGFR?

All ten users answered in the affirmative. The verb "bind" indicates a physical interaction. We used BioGrid to check if there was a protein-protein interaction between TGFA and EGFR, and sure enough BioGrid pointed us to a study that confirmed a physical interaction between those two proteins [7]. We looked at the next sentence:

> ADAMTS20 could be required directly for cleavage of either Kit and/or Kitl to produce sKitl.

Users were asked:

> Does this sentence indicate that cleavage best describes the protein interaction between Kit and Kitl?

For this question, regardless of the true answer, it does not seem that we could conclude that cleavage best describes the interaction between Kit and Kitl based on the sentence alone. Our users split on this question with four voting in the affirmative and six voting against. We were satisfied to see that the majority rejected the interaction.

5. Related & Future Work

5.1 Related Work

Donaldson et al. created a literature-mining system that utilizes Support Vector Machine (SVM) to collect protein–protein interaction information [4]. Once they find an interaction associated with an abstract from PubMed, they ask curators to confirm the connection. Curators are expected to read the abstract or the full article in order to perform such a task. This differs greatly from our approach of showing users only a sentence and asking them to make a determination. This in part is what enables us to utilize non-experts and to attempt to crowd-system our confirmation process.

Marcotte et al. use a Bayesian approach to assign scores to Medline abstracts in order to determine which articles describe interactions between yeast proteins [10]. They do not take the next step and generate a protein–protein interaction database. Instead, similar to the work done by Donaldson et al., a curator could use their output to know which articles to read in order to manually extract the relevant information.

Ono et al. avoid the complexities of natural language processing (NLP) techniques by creating simple rules associated with protein–protein interaction in their attempt to mine literature [11]. When they process sentences, however, they only look for four key words: "interact," "associate," "bind," and "complex." This limited vocabulary, along with very specific sentence structure requirements, makes this algorithm likely to miss interactions and also unable to recognize many important interactions such as ubiquitination and phosphorylation.

Friedman et al. created a NLP system, called GENIE, to extraction molecular pathways from full articles [6]. They attempt to extract many types of relationships, and as such their precision is limited by the type of relationship to be extracted and the literature corpus to be processed. The review process requires an expert to spend several hours per article in order to manually gather relationships, so the authors ended up having to limit their evaluation to just one article.

Blaschke et al. designed a system that detects protein–protein interactions from abstracts [3]. They formed a set of fourteen words associated with protein interaction, and use that along with a set of rules and constraints in order to extract informative sentences. By imposing such limits, they are able to cull out a lot of sentences, but they also invariably miss interactions due to these limitations.

5.2 Future Work

We are still adjusting our ranking algorithm in order to optimize our scoring system. We could potentially ask users to rank the difficulty of sentences and then add a machine learning component. Currently, though, we suspect that we can come up with better metrics through our observations that would result in a more significant increase in performance. There also might be other NLP ideas that we could potentially incorporate to better rank sentences.

We would like to assign initial weights to users depending on how well they perform on known interactions. However, we currently lack pathways that have very specific protein–protein interaction labeling, making such a task difficult. We are hoping to collaborate with a curator in the near-future in order to overcome that obstacle.

Additionally, a curator could help us in the evaluation step by informing us if we were able to discover an interaction currently not in any database.

We would like to establish a point system for users so that they can compete and show off their high scores. Even when we have more data that we can use to evaluate these users, we are still confronted with the fact that they will inevitably be answering questions for which we lack an answer. One potential idea we had was to implement a gambling-like system, where users bet their points on some of their answers. Then after we have expert users answer these questions, we could reward these players. Those players that answered contested questions correctly would gain much more points than those players that answer questions that were nearly unanimous.

We would also like to create a framework that allows two teams to complete against each other. We envision schools or clubs playing against each other while helping us collect data. Our entry into Google Hangout which allows collaborative play was a way to start promoting that idea. Games are not only more fun when accompanied by friends, but we also suspect that their answers would likely be more accurate due to the peer-review nature of playing as a team.

Lastly, we would like to allow our users to target the disease that is most relevant to them. In other words, if a user wants to help fight breast cancer, they could log in to our system and select that as their cause. We would then give them pathways that are the most associated with breast cancer. We also see this as another way to entice user to participate in our game. There will probably be a lot of overlap between pathways and various diseases regardless. Still, we expect users would feel invigorated to help push research in a field that they are passionate about.

6. Conclusion

We have shown a novel method of enabling untrained users without biological backgrounds to successfully participate in generating pathways and protein-protein interactions by reading stand-alone sentences from published biological research. While other proposed methods struggled on recruiting an expert to verify their mined data, our system is able to function almost completely without them. Although we take advantage of experts to affirm contested questions and make substantial pathway modifications, we are still able to function on just the consensus of our regular users.

Making a game that users will enjoy and freely contribute to is particularly challenging when it pertains to a highly scientific field like biology. We have shown with our user feedback, however, that our game is challenging yet satisfying. Users overwhelmingly stated that they would play our game even without a monetary incentive. This is an important threshold for games that rely on crowd-sourcing to drive their data. It is easy to add a few whistles to any system and then call it a game, but if users see the task as poorly disguised work, they will not play it.

We also wanted our users to be able to walk away having learned something about biology. We try to be unobtrusive with the biological information in order to not overload users at the start of the game, but we expect them to eventually explore those hidden options for their betterment. Certainly it benefits us to have a more educated user base, but that is just a side benefit. Instead, when a user contributes to our database, we hope that the user grows as well.

A. Amazon Mechanical Turk User Surveys

We ran two user studies by recruiting users from Amazon Mechanical Turk (MTurk) [1]. We paid them one dollar to perform our task. Users were instructed to go to our web application and answer ten questions. Upon answering those questions, they were given a survey code. After that, they were to return to the MTurk website to input their survey code and answer the following questions:

- What is your opinion on the difficulty of the BioTurk questions?
- Did you enjoy answering the questions?
- Would have you answered these questions for free if you knew that they were used to advanced disease research?
- Any suggestions for us? What is your overall opinion?
- What is your gender?
- What is your age?
- Which country/state are you located in?
- Which of the following best describes your highest achieved education level?
- What is the total income of your household?

We performed one study on March 11, 2013 (Table 1). After reflecting on the feedback and changing the design of BioTurk, we performed another study on May 8, 2013 (Table 2).

Table 1: Amazon Mechanical Turk User Survey – March 11, 2013 – Old Interface

Sex	Age	Loc	Education	Income	Difficulty	Enjoyment	Free?	Suggestion?
F	36	USA	Bachelors degree	$37,500 - $49,999	The questions were difficult.	I enjoyed answering the questions. I also enjoy a task in which I may learn something new.	I would answer a few questions for free in order to help research.	I enjoyed this survey.
F	34	OH	Associates degree	$50,000 - $62,499	extremely difficult.	Nope they were hard for me.	possibly.	Make a background in biology a requirement. Very hard.
F	33	NY	Bachelors degree	$25,000 - $37,499	they were very hard	no	yes	
M	43	TX	Graduate degree	$100,000 or More	they were fairly hard	yes they were fine	i cant say that i would	it is a little difficult for the lay person to understand but i am sure they will try as much as I did to answer the questions
F	57	USA	Some college, no degree	$25,000 - $37,499	Very difficult - I didn't understand them at all.	Sort of. I was trying to puzzle them out, but had no confidence in my understanding.	No, due to my lack of confidence in my understanding of the questions.	A bit too much for the lay person! Someone with training would do a much better job.
M	39	NC	Associates degree	$25,000 - $37,499	Quite difficult, but possible to determine!	Sure!	Probably not	No suggestions, interesting!
F	32	CA	Bachelors degree	$25,000 - $37,499	Very difficult!	Not really, I could barely understand them.	Of course if I helped at all!	I liked the setup! The only thing is there were still more questions after I got my code so I answered another one and got another code–kind of confusing.
F	60	NC	Bachelors degree	$37,500 - $49,999	IT WAS DIFFICULT TO ANALYZE.	NOT REALLY.	NO. ONLY IF THEY WERE EASIER TO COMPREHEND.	THEY SEEMED DIFFICULT TO THE NON SCIENTIFIC MIND.
F	32	WA	Bachelors degree	$25,000 - $37,499	They are very difficult for someone that doesn't understand biology/chemistry. I had some difficulty with it.	Yes, even though I am not sure how many I answered correctly.	No. I need money now.	It was pretty fun. I enjoyed thinking about the answers.
M	38	FL	Bachelors degree	$100,000 or More	they weren't *brutal*, but my knowledge of biochemical reactions is rusty at best	yes, made me think – which is always a good thing	tough to say	kind of liked it
F	51	No data	Some college, no degree	$12,500 - $24,999	They were very difficult for someone who has almost no prior knowledge. I had to concentrate hard and read each sentence and question several times before I reluctantly answered.	No, not really, because I was so unsure about my answers. I didn't feel confident.	No probably not. I'm not sure how my answers would advance disease research? I did my best but I was probably wrong at least some of the time.	The interface was a bit confusing. I stopped answering when a survey code was displayed although there was still the option of answering a question. I hope this was correct. It also took me a while to realize that there are pop up boxes that give further explanations. This could have been explained better at the start. Was it required to leave a comment in the pop up comment box?

Table 2: Amazon Mechanical Turk User Survey – May 8, 2013 – Newer Interface

Sex	Age	Loc	Education	Income	Difficulty	Enjoyment	Free?	Suggestion?
M	50	SC	Some High School	$25,000 - $37,499	I found them to be quite difficult. Reading the descriptions and the question several times was necessary for me to try and complete the task.	Actually I did. I would like to know how I did.	Perhaps. Not on the mechanical turk interface.	Well designed for the most part. Some of the additional text and descriptions are too small to read without zoom.
F	23	USA	Bachelors degree	$62,500 - $74,999	They were rather difficult.	They were okay.	Yes. :)	Very interesting, although hard.
M	44	FL	Bachelors degree	$50,000 - $62,499	The questions were very difficult and specialized.	Not really. I did enjoy the interface.	Yes, I would for free. If I had the proper knowledge to answer them correctly.	No suggestions. I thought the continuous movement was a little bit unnecessary.
F	66	USA	Graduate degree	No data	very difficult	Not really because I did not understand them. I just hadto go with sentence structure	yes	interesting–if for a good reason
F	33	NY	Bachelors degree	$37,500 - $49,999	they were very hard for me to understand	no not really	maybe	I liked the interface and the colors of the circles
M	48	USA	Some college, no degree	$12,500 - $24,999	Extremely difficult	Did not understand about 90% of the questions	not sure	The site is easy to use
M	39	NC	Associates degree	$37,500 - $49,999	The wording is far out of my known, so that was a bit tough.	Sure, it wasnt exciting, but it wasnt bad either.	Possibly depending on the situation I was asked in.	It was straight forward, no issues!
M	43	NC	Some college, no degree	$25,000 - $37,499	The biology terms are pretty far above my head, but I think I did ok with comparing the diagrams and the wording of the sentences.	Yes, it was different than most mTurk tasks.	No	Maybe in the tutorial, give some examples of what the user should be looking for and how to compare the wording to the diagram.
F	61	GA	Bachelors degree	$62,500 - $74,999	EXTREMELY	no	yes, of course	i have no suggestions, because i don't understand the purpose
M	30	IL	High School Graduate	$25,000 - $37,499	It was difficult as I don't have any knowledge of the field. And without explanatory sample questions or a visible grading system, I lost any modicum of confidence I had and quickly found myself second guessing all of my answers.	It was OK. The interface was aesthetically pleasing and easy to use.	No.	Provide a tutorial covering the subject and include a selection of sample questions that explain the reasoning behind the correct answer.

Bibliography

[1] Amazon Mechanical Turk. `https://www.mturk.com/mturk/welcome`, March 2013.

[2] Amos Bairoch, Rolf Apweiler, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. The universal protein resource (uniprot). *Nucleic acids research*, 33(suppl 1):D154–D159, 2005.

[3] Christian Blaschke, Miguel A Andrade, Christos Ouzounis, and Alfonso Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proc Int Conf Intell Syst Mol Biol*, volume 7, pages 60–67, 1999.

[4] Ian Donaldson, Joel Martin, Berry de Bruijn, Cheryl Wolting, Vicki Lay, Brigitte Tuekam, Shudong Zhang, Berivan Baskin, Gary D Bader, Katerina Michalickova, et al. Prebind and textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC bioinformatics*, 4(1):11, 2003.

[5] Ran Elkon, Rita Vesterman, Nira Amit, Igor Ulitsky, Idan Zohar, Mali Weisz, Gilad Mass, Nir Orlev, Giora Sternberg, Ran Blekhman, et al. Spike–a database, visualization and analysis tool of cellular signaling pathways. *BMC bioinformatics*, 9(1):110, 2008.

[6] Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(suppl 1):S74–S82, 2001.

[7] Thomas PJ Garrett, Neil M McKern, Meizhen Lou, Thomas C Elleman, Timothy E Adams, George O Lovrecz, Hong-Jian Zhu, Francesca Walker, Morry J Frenkel, Peter A Hoyne, et al. Crystal structure of a truncated epidermal growth factor receptor extracellular domain bound to transforming growth factor α. *Cell*, 110(6):763–773, 2002.

[8] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428–D432, 2005.

[9] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The kegg resource for deciphering the genome. *Nucleic acids research*, 32(suppl 1):D277–D280, 2004.

[10] Edward M Marcotte, Ioannis Xenarios, and David Eisenberg. Mining literature for protein–protein interactions. *Bioinformatics*, 17(4):359–363, 2001.

[11] Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami, and Toshihisa Takagi. Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161, 2001.

[12] Philipp Pagel, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stümpflen, Hans-Werner Mewes, et al. The mips mammalian protein–protein interaction database. *Bioinformatics*, 21(6):832–834, 2005.

[13] Petra Paul, Tineke van den Hoorn, Marlieke LM Jongsma, Mark J Bakker, Rutger Hengeveld, Lennert Janssen, Peter Cresswell, David A Egan, Marieke van Ham, Anja ten Brinke, et al. A genome-wide multidimensional rnai screen reveals pathways controlling mhc class ii antigen presentation. *Cell*, 145(2):268–283, 2011.

[14] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl 1):D674–D679, 2009.

[15] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[16] Andrew Su. Gene Games. `http://genegames.org/games/`, April 2013.

[17] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.