# Expert Search - Automated Crawler

Runs to discover Faculty Bio-pages by crawling University Websites.

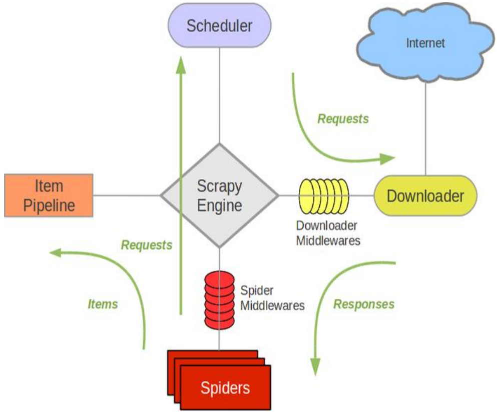
Mohana Venkata Kalyan Cheerla

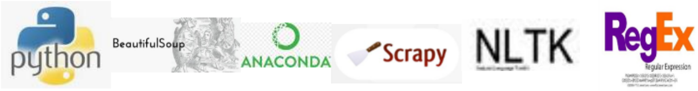
# Why an Automated Crawler?

- ➔ Manually searching the web and extracting faculty bio pages are time consuming
  - ◆ It's tedious
  - ◆ Doesn't scale
  - ◆ Navigation is difficult
- ➔ Our Automated Crawler crawls university websites to automatically fetches bio pages
- ➔ It converts the Markup text into normalized text data to feed its downstream components

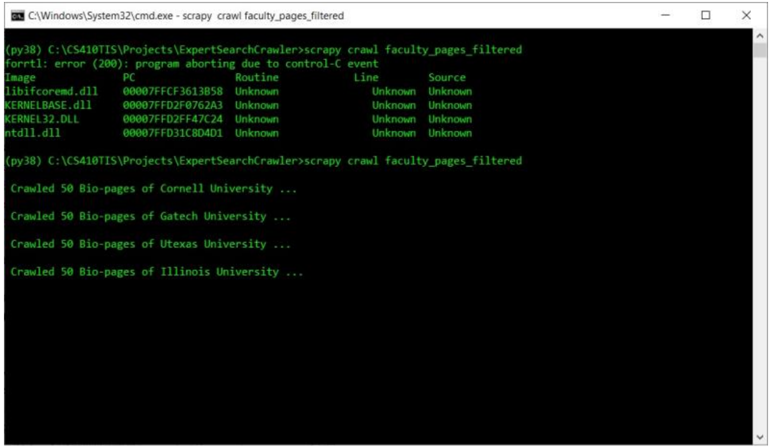
# Technology Stack

- ➔ The crawler runs on a cron schedule to to passively crawl universities websites
- ➔ The crawler is developed in Python. It utilizes the Scrapy library and follows Breadth First Search (BFS) up to the designated depth specified by the User/Administrator
- ➔ Crawler has Spiders that recursively follow all the web links of the University Website and extracts the web page data along with the Website's "URL Tree" to understand, validate, troubleshoot and trace



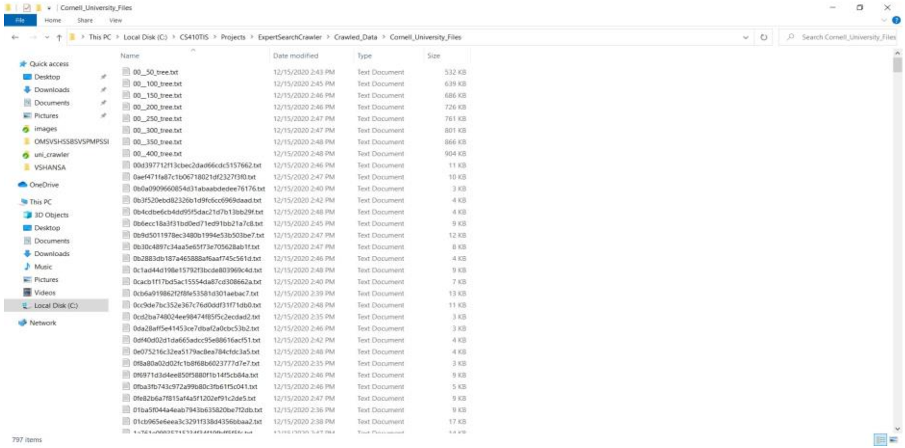


# Crawler in Action



C:\Windows\System32\cmd.exe - scrapy  crawl faculty_pages_filtered

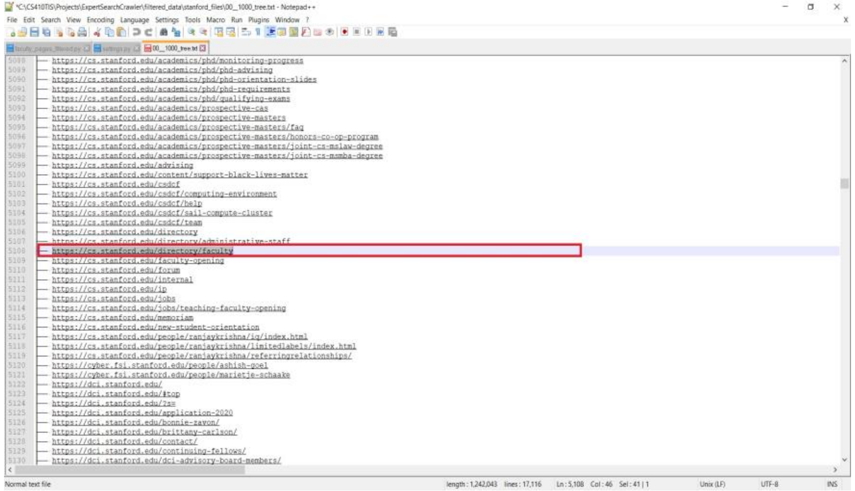
```
(py38) C:\CS410TIS\Projects\ExpertSearchCrawler>scrapy crawl faculty_pages_filtered
forrtl: error (200): program aborting due to control-C event
Image              PC                Routine            Line        Source
libifcoremd.dll    00007FFCF3613B58  Unknown               Unknown  Unknown
KERNELBASE.dll     00007FFD2F0762A3  Unknown               Unknown  Unknown
KERNEL32.DLL       00007FFD2FF47C24  Unknown               Unknown  Unknown
ntdll.dll          00007FFD31C8D4D1  Unknown               Unknown  Unknown

(py38) C:\CS410TIS\Projects\ExpertSearchCrawler>scrapy crawl faculty_pages_filtered

 Crawled 50 Bio-pages of Cornell University ...

 Crawled 50 Bio-pages of Gatech University ...

 Crawled 50 Bio-pages of Utexas University ...

 Crawled 50 Bio-pages of Illinois University ...
```



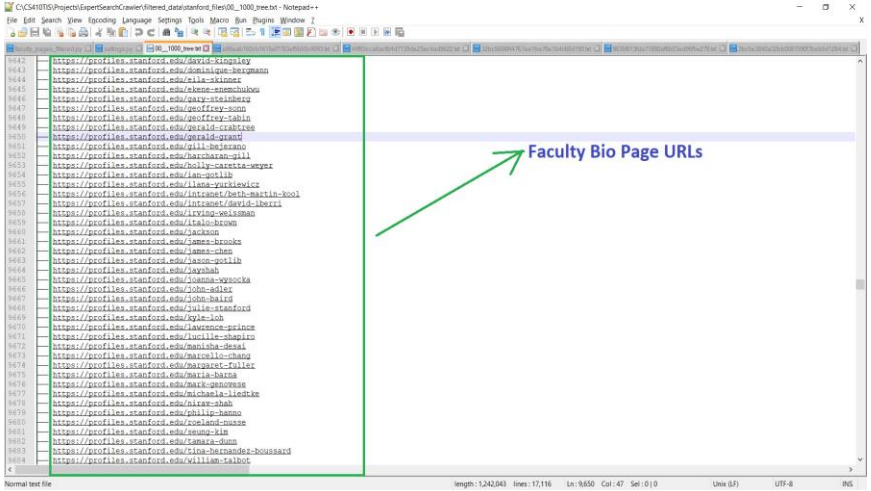
| Name | Date modified | Type | Size |
|---|---|---|---|
| 00__50_tree.txt | 12/15/2020 2:43 PM | Text Document | 532 KB |
| 00__100_tree.txt | 12/15/2020 2:45 PM | Text Document | 639 KB |
| 00__150_tree.txt | 12/15/2020 2:46 PM | Text Document | 686 KB |
| 00__200_tree.txt | 12/15/2020 2:46 PM | Text Document | 726 KB |
| 00__250_tree.txt | 12/15/2020 2:47 PM | Text Document | 761 KB |
| 00__300_tree.txt | 12/15/2020 2:47 PM | Text Document | 801 KB |
| 00__350_tree.txt | 12/15/2020 2:48 PM | Text Document | 866 KB |
| 00__400_tree.txt | 12/15/2020 2:48 PM | Text Document | 904 KB |
| 00d397712f13cbec2dad66cdc5157662.txt | 12/15/2020 2:46 PM | Text Document | 11 KB |
| 0aef471fa87c1b06718021df2327f3f0.txt | 12/15/2020 2:47 PM | Text Document | 10 KB |
| 0b0a0909660854d31abaabdedee76176.txt | 12/15/2020 2:40 PM | Text Document | 3 KB |
| 0b3f520ebd82326b1d9fc6cc6969daad.txt | 12/15/2020 2:42 PM | Text Document | 4 KB |
| 0b4cdbe6cb4dd95f5dac21d7b13bb29f.txt | 12/15/2020 2:48 PM | Text Document | 4 KB |
| 0b6ecc18a3f31bd0ed71ed91bb21a7c8.txt | 12/15/2020 2:45 PM | Text Document | 9 KB |
| 0b9d5011978ec3480b1994e53b503be7.txt | 12/15/2020 2:47 PM | Text Document | 12 KB |
| 0b30c4897c34aa5e65f73e705628ab1f.txt | 12/15/2020 2:47 PM | Text Document | 8 KB |
| 0b2883db187a465888af6aaf745c561d.txt | 12/15/2020 2:46 PM | Text Document | 4 KB |
| 0c1ad44d198e15792f3bcde803969c4d.txt | 12/15/2020 2:48 PM | Text Document | 9 KB |
| 0cacb1f17bd5ac15554da87cd308662a.txt | 12/15/2020 2:40 PM | Text Document | 7 KB |
| 0cb6a919862f2f8fe53581d301aebac7.txt | 12/15/2020 2:39 PM | Text Document | 13 KB |
| 0cc9de7bc352e367c76d0ddf31f71db0.txt | 12/15/2020 2:48 PM | Text Document | 11 KB |
| 0cd2ba748024ee98474f85f5c2ecdad2.txt | 12/15/2020 2:35 PM | Text Document | 3 KB |
| 0da28aff5e41453ce7dbaf2a0cbc53b2.txt | 12/15/2020 2:46 PM | Text Document | 3 KB |
| 0df40d02d1da665adcc95e88616acf51.txt | 12/15/2020 2:42 PM | Text Document | 4 KB |
| 0e075216c32ea5179ac8ea784cfdc3a5.txt | 12/15/2020 2:48 PM | Text Document | 4 KB |
| 0f8a80a02d02fc1b8f68b6023777d7e7.txt | 12/15/2020 2:35 PM | Text Document | 3 KB |
| 0f6971d3d4ee850f5880f1b14f5cb84a.txt | 12/15/2020 2:46 PM | Text Document | 9 KB |
| 0fba3fb743c972a99b80c3fb61f5c041.txt | 12/15/2020 2:46 PM | Text Document | 5 KB |
| 0fe82b6a7f815af4a5f1202ef91c2de5.txt | 12/15/2020 2:47 PM | Text Document | 9 KB |
| 01ba5f044a4eab7943b635820be7f2db.txt | 12/15/2020 2:36 PM | Text Document | 9 KB |
| 01cb965e6eea3c3291f338d4356bbaa2.txt | 12/15/2020 2:38 PM | Text Document | 17 KB |

797 items

# Crawler in Action (contd…)


*C:\CS410TIS\Projects\ExpertSearchCrawler\filtered_data\stanford_files\00__1000_tree.txt - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
00__1000_tree.txt
Normal text file
length : 1,242,043 lines : 17,116 Ln : 5,108 Col : 46 Sel : 41 | 1 Unix (LF) UTF-8 INS

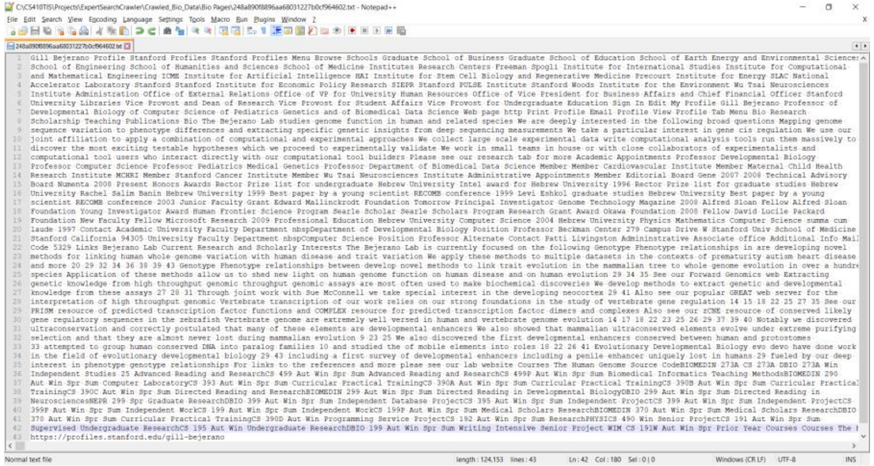
```
├── https://cs.stanford.edu/academics/phd/monitoring-progress
├── https://cs.stanford.edu/academics/phd/phd-advising
├── https://cs.stanford.edu/academics/phd/phd-orientation-slides
├── https://cs.stanford.edu/academics/phd/phd-requirements
├── https://cs.stanford.edu/academics/phd/qualifying-exams
├── https://cs.stanford.edu/academics/prospective-cas
├── https://cs.stanford.edu/academics/prospective-masters
├── https://cs.stanford.edu/academics/prospective-masters/faq
├── https://cs.stanford.edu/academics/prospective-masters/honors-co-op-program
├── https://cs.stanford.edu/academics/prospective-masters/joint-cs-mslaw-degree
├── https://cs.stanford.edu/academics/prospective-masters/joint-cs-msmba-degree
├── https://cs.stanford.edu/advising
├── https://cs.stanford.edu/content/support-black-lives-matter
├── https://cs.stanford.edu/csdcf
├── https://cs.stanford.edu/csdcf/computing-environment
├── https://cs.stanford.edu/csdcf/help
├── https://cs.stanford.edu/csdcf/sail-compute-cluster
├── https://cs.stanford.edu/csdcf/team
├── https://cs.stanford.edu/directory
├── https://cs.stanford.edu/directory/administrative-staff
├── https://cs.stanford.edu/directory/faculty
├── https://cs.stanford.edu/faculty-opening
├── https://cs.stanford.edu/forum
├── https://cs.stanford.edu/internal
├── https://cs.stanford.edu/ip
├── https://cs.stanford.edu/jobs
├── https://cs.stanford.edu/jobs/teaching-faculty-opening
├── https://cs.stanford.edu/memoriam
├── https://cs.stanford.edu/new-student-orientation
├── https://cs.stanford.edu/people/ranjaykrishna/iq/index.html
├── https://cs.stanford.edu/people/ranjaykrishna/limitedlabels/index.html
├── https://cs.stanford.edu/people/ranjaykrishna/referringrelationships/
├── https://cyber.fsi.stanford.edu/people/ashish-goel
├── https://cyber.fsi.stanford.edu/people/marietje-schaake
├── https://dci.stanford.edu/
├── https://dci.stanford.edu/#top
├── https://dci.stanford.edu/?s=
├── https://dci.stanford.edu/application-2020
├── https://dci.stanford.edu/bonnie-zavon/
├── https://dci.stanford.edu/brittany-carlson/
├── https://dci.stanford.edu/contact/
├── https://dci.stanford.edu/continuing-fellows/
├── https://dci.stanford.edu/dci-advisory-board-members/
```


C:\CS410TIS\Projects\ExpertSearchCrawler\filtered_data\stanford_files\00__1000_tree.txt - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
00__1000_tree.txt
Normal text file
length : 1,242,043 lines : 17,116 Ln : 9,650 Col : 47 Sel : 0 | 0 Unix (LF) UTF-8 INS

```
├── https://profiles.stanford.edu/david-kingsley
├── https://profiles.stanford.edu/dominique-bergmann
├── https://profiles.stanford.edu/eila-skinner
├── https://profiles.stanford.edu/ekene-enemchukwu
├── https://profiles.stanford.edu/gary-steinberg
├── https://profiles.stanford.edu/geoffrey-sonn
├── https://profiles.stanford.edu/geoffrey-tabin
├── https://profiles.stanford.edu/gerald-crabtree
├── https://profiles.stanford.edu/gerald-grant
├── https://profiles.stanford.edu/gill-bejerano
├── https://profiles.stanford.edu/harcharan-gill
├── https://profiles.stanford.edu/holly-caretta-weyer
├── https://profiles.stanford.edu/ian-gotlib
├── https://profiles.stanford.edu/ilana-yurkiewicz
├── https://profiles.stanford.edu/intranet/beth-martin-kool
├── https://profiles.stanford.edu/intranet/david-iberri
├── https://profiles.stanford.edu/irving-weissman
├── https://profiles.stanford.edu/italo-brown
├── https://profiles.stanford.edu/jackson
├── https://profiles.stanford.edu/james-brooks
├── https://profiles.stanford.edu/james-chen
├── https://profiles.stanford.edu/jason-gotlib
├── https://profiles.stanford.edu/jayshah
├── https://profiles.stanford.edu/joanna-wysocka
├── https://profiles.stanford.edu/john-adler
├── https://profiles.stanford.edu/john-baird
├── https://profiles.stanford.edu/julie-stanford
├── https://profiles.stanford.edu/kyle-loh
├── https://profiles.stanford.edu/lawrence-prince
├── https://profiles.stanford.edu/lucille-shapiro
├── https://profiles.stanford.edu/manisha-desai
├── https://profiles.stanford.edu/marcello-chang
├── https://profiles.stanford.edu/margaret-fuller
├── https://profiles.stanford.edu/maria-barna
├── https://profiles.stanford.edu/mark-genovese
├── https://profiles.stanford.edu/michaela-liedtke
├── https://profiles.stanford.edu/nirav-shah
├── https://profiles.stanford.edu/philip-hanno
├── https://profiles.stanford.edu/roeland-nusse
├── https://profiles.stanford.edu/seung-kim
├── https://profiles.stanford.edu/tamara-dunn
├── https://profiles.stanford.edu/tina-hernandez-boussard
├── https://profiles.stanford.edu/william-talbot
```

Faculty Bio Page URLs

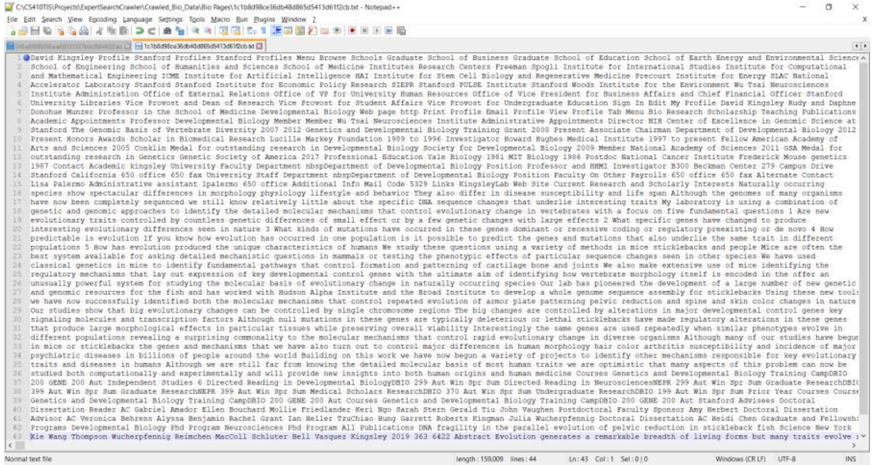
# Crawler in Action (contd...)



File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
248a890f8896aa68031227b0cf964602.txt
Normal text file
length : 124,153 lines : 43
Ln : 42 Col : 180 Sel : 0 | 0
Windows (CR LF) UTF-8 INS

C:\CS410TIS\Projects\ExpertSearchCrawler\Crawled_Bio_Data\Bio Pages\248a890f8896aa68031227b0cf964602.txt - Notepad++

```
Gill Bejerano Profile Stanford Profiles Stanford Profiles Menu Browse Schools Graduate School of Business Graduate School of Education School of Earth Energy and Environmental Science:
School of Engineering School of Humanities and Sciences School of Medicine Institutes Research Centers Freeman Spogli Institute for International Studies Institute for Computational
and Mathematical Engineering ICME Institute for Artificial Intelligence HAI Institute for Stem Cell Biology and Regenerative Medicine Precourt Institute for Energy SLAC National
Accelerator Laboratory Stanford Stanford Institute for Economic Policy Research SIEPR Stanford PULSE Institute Stanford Woods Institute for the Environment Wu Tsai Neurosciences
Institute Administration Office of External Relations Office of VP for University Human Resources Office of Vice President for Business Affairs and Chief Financial Officer Stanford
University Libraries Vice Provost and Dean of Research Vice Provost for Student Affairs Vice Provost for Undergraduate Education Sign In Edit My Profile Gill Bejerano Professor of
Developmental Biology of Computer Science of Pediatrics Genetics and of Biomedical Data Science Web page http Print Profile Email Profile View Profile Tab Menu Bio Research
Scholarship Teaching Publications Bio The Bejerano Lab studies genome function in human and related species We are deeply interested in the following broad questions Mapping genome
sequence variation to phenotype differences and extracting specific genetic insights from deep sequencing measurements We take a particular interest in gene cis regulation We use our
joint affiliation to apply a combination of computational and experimental approaches We collect large scale experimental data write computational analysis tools run them massively to
discover the most exciting testable hypotheses which we proceed to experimentally validate We work in small teams in house or with close collaborators of experimentalists and
computational tool users who interact directly with our computational tool builders Please see our research tab for more Academic Appointments Professor Developmental Biology
Professor Computer Science Professor Pediatrics Medical Genetics Professor Department of Biomedical Data Science Member Member Cardiovascular Institute Member Maternal Child Health
Research Institute MCHRI Member Stanford Cancer Institute Member Wu Tsai Neurosciences Institute Administrative Appointments Member Editorial Board Gene 2007 2008 Technical Advisory
Board Numenta 2008 Present Honors Awards Rector Prize list for undergraduate Hebrew University Intel award for Hebrew University 1996 Rector Prize list for graduate studies Hebrew
University Rachel Salim Banin Hebrew University 1999 Best paper by a young scientist RECOMB conference 1999 Levi Eshkol graduate studies Hebrew University Best paper by a young
scientist RECOMB conference 2003 Junior Faculty Grant Edward Mallinckrodt Foundation Tomorrow Principal Investigator Genome Technology Magazine 2008 Alfred Sloan Fellow Alfred Sloan
Foundation Young Investigator Award Human Frontier Science Program Searle Scholar Searle Scholars Program Research Grant Award Okawa Foundation 2008 Fellow David Lucile Packard
Foundation New Faculty Fellow Microsoft Research 2009 Professional Education Hebrew University Computer Science 2004 Hebrew University Physics Mathematics Computer Science summa cum
laude 1997 Contact Academic University Faculty Department nbspDepartment of Developmental Biology Position Professor Beckman Center 279 Campus Drive W Stanford Univ School of Medicine
Stanford California 94305 University Faculty Department nbspComputer Science Position Professor Alternate Contact Patti Livingston Admininstrative Associate office Additional Info Mail
Code 5329 Links Bejerano Lab Current Research and Scholarly Interests The Bejerano Lab is currently focused on the following Genotype Phenotype relationships in are developing novel
methods for linking human whole genome variation with human disease and trait variation We apply these methods to multiple datasets in the contexts of prematurity autism heart disease
and more 20 29 32 34 36 38 39 43 Genotype Phenotype relationships between develop novel methods to link trait evolution in the mammalian tree to whole genome evolution in over a hundre
species Application of these methods allow us to shed new light on human genome function on human disease and on human evolution 29 34 35 See our Forward Genomics web Extracting
genetic knowledge from high throughput genomic throughput genomic assays are most often used to make biochemical discoveries We develop methods to extract genetic and developmental
knowledge from these assays 27 28 31 Through joint work with Sue McConnell we take special interest in the developing neocortex 29 41 Also see our popular GREAT web server for the
interpretation of high throughput genomic Vertebrate transcription of our work relies on our strong foundations in the study of vertebrate gene regulation 14 15 18 22 25 27 35 See our
PRISM resource of predicted transcription factor functions and COMPLEX resource for predicted transcription factor dimers and complexes Also see our zCNE resource of conserved likely
gene regulatory sequences in the zebrafish Vertebrate genome are extremely well versed in human and vertebrate genome evolution 14 17 18 22 23 25 26 29 37 39 40 Notably we discovered
ultraconservation and correctly postulated that many of these elements are developmental enhancers We also showed that mammalian ultraconserved elements evolve under extreme purifying
selection and that they are almost never lost during mammalian evolution 9 23 25 We also discovered the first developmental enhancers conserved between human and protostomes
33 attempted to group human conserved DNA into paralog families 10 and studied the of mobile elements into roles 18 22 26 41 Evolutionary Developmental Biology evo devo have done work
in the field of evolutionary developmental biology 29 43 including a first survey of developmental enhancers including a penile enhancer uniquely lost in humans 29 fueled by our deep
interest in phenotype genotype relationships For links to the references and more please see our lab website Courses The Human Genome Source CodeBIOMEDIN 273A CS 273A DBIO 273A Win
Independent Studies 25 Advanced Reading and ResearchCS 499 Aut Win Spr Sum Advanced Reading and ResearchCS 499P Aut Win Spr Sum Biomedical Informatics Teaching MethodsBIOMEDIN 290
Aut Win Spr Sum Computer LaboratoryCS 393 Aut Win Spr Sum Curricular Practical TrainingCS 390A Aut Win Spr Sum Curricular Practical TrainingCS 390B Aut Win Spr Sum Curricular Practical
TrainingCS 390C Aut Win Spr Sum Directed Reading and ResearchBIOMEDIN 299 Aut Win Spr Sum Directed Reading in Developmental BiologyDBIO 299 Aut Win Spr Sum Directed Reading in
NeurosciencesNEPR 299 Spr Graduate ResearchDBIO 399 Aut Win Spr Sum Independent Database ProjectCS 395 Aut Win Spr Sum Independent ProjectCS 399 Aut Win Spr Sum Independent ProjectCS
399P Aut Win Spr Sum Independent WorkCS 199 Aut Win Spr Sum Independent WorkCS 199P Aut Win Spr Sum Medical Scholars ResearchBIOMEDIN 370 Aut Win Spr Sum Medical Scholars ResearchDBIO
370 Aut Win Spr Sum Curricular Practical TrainingCS 390D Aut Win Programming Service ProjectCS 192 Aut Win Spr Sum ResearchPHYSICS 490 Win Senior ProjectCS 191 Aut Win Spr Sum
Supervised Undergraduate ResearchCS 195 Aut Win Undergraduate ResearchDBIO 199 Aut Win Spr Sum Writing Intensive Senior Project WIM CS 191W Aut Win Spr Prior Year Courses Courses The 
https://profiles.stanford.edu/gill-bejerano
```



File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
1c1b8d98ce36db48d865d5413d61f2cb.txt
Normal text file
length : 159,009 lines : 44
Ln : 43 Col : 1 Sel : 0 | 0
Windows (CR LF) UTF-8 INS

C:\CS410TIS\Projects\ExpertSearchCrawler\Crawled_Bio_Data\Bio Pages\1c1b8d98ce36db48d865d5413d61f2cb.txt - Notepad++

```
David Kingsley Profile Stanford Profiles Stanford Profiles Menu Browse Schools Graduate School of Business Graduate School of Education School of Earth Energy and Environmental Science
School of Engineering School of Humanities and Sciences School of Medicine Institutes Research Centers Freeman Spogli Institute for International Studies Institute for Computational
and Mathematical Engineering ICME Institute for Artificial Intelligence HAI Institute for Stem Cell Biology and Regenerative Medicine Precourt Institute for Energy SLAC National
Accelerator Laboratory Stanford Stanford Institute for Economic Policy Research SIEPR Stanford PULSE Institute Stanford Woods Institute for the Environment Wu Tsai Neurosciences
Institute Administration Office of External Relations Office of VP for University Human Resources Office of Vice President for Business Affairs and Chief Financial Officer Stanford
University Libraries Vice Provost and Dean of Research Vice Provost for Student Affairs Vice Provost for Undergraduate Education Sign In Edit My Profile David Kingsley Rudy and Daphne
Donohue Munzer Professor in the School of Medicine Developmental Biology Web page http Print Profile Email Profile View Profile Tab Menu Bio Research Scholarship Teaching Publications
Academic Appointments Professor Developmental Biology Member Member Wu Tsai Neurosciences Institute Administrative Appointments Director NIH Center of Excellence in Genomic Science at
Stanford The Genomic Basis of Vertebrate Diversity 2007 2012 Genetics and Developmental Biology Training Grant 2008 Present Associate Chairman Department of Developmental Biology 2012
Present Honors Awards Scholar in Biomedical Research Lucille Markey Foundation 1989 to 1996 Investigator Howard Hughes Medical Institute 1997 to present Fellow American Academy of
Arts and Sciences 2005 Conklin Medal for outstanding research in Developmental Biology Society for Developmental Biology 2009 Member National Academy of Sciences 2011 GSA Medal for
outstanding research in Genetics Genetic Society of America 2017 Professional Education Yale Biology 1981 MIT Biology 1986 Postdoc National Cancer Institute Frederick Mouse genetics
1987 Contact Academic kingsley University Faculty Department nbspDepartment of Developmental Biology Position Professor and HHMI Investigator B300 Beckman Center 279 Campus Drive
Stanford California 650 office 650 fax University Staff Department nbspDepartment of Developmental Biology Position Faculty On Other Payrolls 650 office 650 fax Alternate Contact
Lisa Palermo Administrative assistant lpalermo 650 office Additional Info Mail Code 5329 Links KingsleyLab Web Site Current Research and Scholarly Interests Naturally occurring
species show spectacular differences in morphology physiology lifestyle and behavior They also differ in disease susceptibility and life span Although the genomes of many organisms
have now been completely sequenced we still know relatively little about the specific DNA sequence changes that underlie interesting traits My laboratory is using a combination of
genetic and genomic approaches to identify the detailed molecular mechanisms that control evolutionary change in vertebrates with a focus on five fundamental questions 1 Are new
evolutionary traits controlled by countless genetic differences of small effect or by a few genetic changes with large effects 2 What specific genes have changed to produce
interesting evolutionary differences seen in nature 3 What kinds of mutations have occurred in these genes dominant or recessive coding or regulatory preexisting or de novo 4 How
predictable is evolution If you know how evolution has occurred in one population is it possible to predict the genes and mutations that also underlie the same trait in different
populations 5 How has evolution produced the unique characteristics of humans We study these questions using a variety of methods in mice sticklebacks and people Mice are often the
best system available for asking detailed mechanistic questions in mammals or testing the phenotypic effects of particular sequence changes seen in other species We have used
classical genetics in mice to identify fundamental pathways that control formation and patterning of cartilage bone and joints We also make extensive use of mice identifying the
regulatory mechanisms that lay out expression of key developmental control genes with the ultimate aim of identifying how vertebrate morphology itself is encoded in the offer an
unusually powerful system for studying the molecular basis of evolutionary change in naturally occurring species Our lab has pioneered the development of a large number of new genetic
and genomic resources for the fish and has worked with Hudson Alpha Institute and the Broad Institute to develop a whole genome sequence assembly for sticklebacks Using these new tools
we have now successfully identified both the molecular mechanisms that control repeated evolution of armor plate patterning pelvic reduction and spine and skin color changes in nature
Our studies show that big evolutionary changes can be controlled by single chromosome regions The big changes are controlled by alterations in major developmental control genes key
signaling molecules and transcription factors Although null mutations in these genes are typically deleterious or lethal sticklebacks have made regulatory alterations in these genes
that produce large morphological effects in particular tissues while preserving overall viability Interestingly the same genes are used repeatedly when similar phenotypes evolve in
different populations revealing a surprising commonality to the molecular mechanisms that control rapid evolutionary change in diverse organisms Although many of our studies have begur
in mice or sticklebacks the genes and mechanisms that we have also turn out to control major differences in human morphology hair color arthritis susceptibility and incidence of major
psychiatric diseases in billions of people around the world Building on this work we have now begun a variety of projects to identify other mechanisms responsible for key evolutionary
traits and diseases in humans Although we are still far from knowing the detailed molecular basis of most human traits we are optimistic that many aspects of this problem can now be
studied both computationally and experimentally and will provide new insights into both human origins and human medicine Courses Genetics and Developmental Biology Training CampDBIO
200 GENE 200 Aut Independent Studies 6 Directed Reading in Developmental BiologyDBIO 299 Aut Win Spr Sum Directed Reading in NeurosciencesNEPR 299 Aut Win Spr Sum Graduate ResearchDBIC
399 Aut Win Spr Sum Graduate ResearchNEPR 399 Aut Win Spr Sum Medical Scholars ResearchDBIO 370 Aut Win Spr Sum Undergraduate ResearchDBIO 199 Aut Win Spr Sum Prior Year Courses Course
Genetics and Developmental Biology Training CampDBIO 200 GENE 200 Aut Courses Genetics and Developmental Biology Training CampDBIO 200 GENE 200 Aut Stanford Advisees Doctoral
Dissertation Reader AC Gabriel Amador Ellen Bouchard Mollie Friedlander Keri Ngo Sarah Stern Gerald Tiu John Vaughen Postdoctoral Faculty Sponsor Amy Herbert Doctoral Dissertation
Advisor AC Veronica Behrens Alyssa Benjamin Rachel Grant Ian Heller TzuChiao Hung Garrett Roberts Kingman Julia Wucherpfennig Doctoral Dissertation AC Heidi Chen Graduate and Fellowshi
Programs Developmental Biology Phd Program Neurosciences Phd Program All Publications DNA fragility in the parallel evolution of pelvic reduction in stickleback fish Science New York
Xie Wang Thompson Wucherpfennig Reimchen MacColl Schluter Bell Vasquez Kingsley 2019 363 6422 Abstract Evolution generates a remarkable breadth of living forms but many traits evolve i
```

# Challenges

- ➔ Crawler runs a long time and drains computational resources on the developer's machine when we run it for all Universities through all the web pages of the university websites without any limitation on its depth.
  - ◆ Parallelization over multiple machines works best
- ➔ Optimizing the crawler was challenging
  - ◆ We had to eliminate different branches of crawl tree that were uninteresting, e.g. publications, news, sports

# Questions?

Please reach out to cheerla3@illinois.edu if you have any question.

# Thank you!!