

SIXING HUANG

BIOINFORMATIC DATA SCIENTIST

CONTACT

4915782039943

dgg321982@gmail.com

Hohes Feld 4, 38124
Braunschweig, Germany

www.linkedin.com/in/sixing-huang-3a824a66

<https://www.dsmz.de/research/bioinformatics/biodiversity-informatics>

OPEN SOURCE PROJECTS

Pyphy - library that interacts with NCBI taxonomy

<https://github.com/dgg32/pyphy>

Meta_detelectron2 - Photo object recognition and indexing

https://github.com/dgg32/meta_detelectron2

Pacbio_its_sequencing

https://github.com/dgg32/pacbio_its_sequencing

LANGUAGES

Chinese | Native

English | Fluent

German | Fluent

Japanese | JLPT N3

SUMMARY

Bioinformatic data scientist eager to contribute to team success with critical thinking, expert coding skill and four languages. Proficient in test-driven development and GitHub and trained in DevOps, machine, deep and reinforcement learning. Graduated as a lab microbiologist and work in DSMZ Germany. Ready to relocate to other parts of the world to begin a new professional journey in biopharmacy, IT, agriculture and industry.

EDUCATION

Ph.D.: Bioinformatics, Biology

Max Planck Institute For Marine Microbiology, Bremen, Germany | 2009-2013

Master of Science: Biology

University Bremen, Bremen, Germany | 2004-2009

EXPERIENCE

Biodiversity Informatics Senior Scientist

Leibniz Institute DSMZ-German Collection Of Microorganisms And Cell Cultures GmbH | Braunschweig, Germany | 2013 - PRESENT

Scientific research with a plethora of techniques: Jupyter, web scraping, data visualization, natural language processing.

- Biodiversity analysis based on NGS 16S rRNA sequences
- Functional single-cell genomics, genomics and metagenomics
- Statistical analysis and machine learning on genomic metadata

Bioinformatics Intern

Max Planck Institute For Marine Microbiology | Bremen, Germany | 2013 - 2013

Analysis of metagenomic data in marine habitats.

- Metagenomic binnings on 454 sequencing data
- Functional genomics of Flavobacteria
- Carbohydrate-degrading enzymes analysis and PUL characterization

SIXING HUANG

BIOINFORMATIC DATA SCIENTIST

CERTIFICATIONS

AWS Fundamentals Specialization

Coursera | 2020

Certified Kubernetes

Administrator (CKA) with Practice Tests

Udemy | 2020

Introduction to Deep Learning

Coursera | 2020

Natural Language Processing

Coursera | 2020

How to Win a Data Science

Competition: Learn from Top Kagglers

Coursera | 2019

Practical Reinforcement Learning

Coursera | 2020

Alibaba Cloud Certified Professional (ACP) Cloud Computing

Udemy | 2020

SKILLS

Python and C++ ★★★★★

MySQL and MongoDB ★★★★★

Machine learning and Deep learning ★★★★★

Hadoop, Spark, Kubernetes and Docker ★★★★★

D3.js, Matplotlib and Inkscape ★★★★★

Slurm, Airflow and Snakemake ★★★★★

AWS, GCP, Azure and Alibaba Cloud ★★★

NGS genomics and metagenomics ★★★★★

PUBLICATION HIGHLIGHTS

- Rohden, F., **Huang, S.**, et al. (2020) Studies on Digital Sequence Information on Genetic Resources (Convention on Biological Diversity commissioned study)
- Dedysh, S., Henke, P. ... **Huang, S.**, et al. (2020) 100-year-old Enigma Solved: Identification, Genomic Characterization and Biogeography of the Yet Uncultured Planctomyces Bekefii. Environ Microbiol. 22(1):198-211.
- Overmann, J., **Huang, S.**, et al. (2019) Relevance of Phenotypic Information for the Taxonomy of Not-Yet-Cultured Microorganisms. Syst Appl Microbiol. 42(1):22-29.
- **Huang, S.**, Vieira, S. et al. (2016) First Complete Genome Sequence of a Subdivision 6 Acidobacterium Strain. Genome Announcements, 4(3):2006–2007.
- Teeling, H., Fuchs, B., ... **Huang, S.**, et al. (2012). Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. Science, 336(6081): 608-11.

SIXING HUANG

BIOINFORMATIC DATA SCIENTIST

ONLINE

Kaggle profile

<https://www.kaggle.com/dgg321982> (TMDB Box Office Prediction: top 43%; IEEE-CIS Fraud Detection: top 45%)

Stack Overflow

<https://stackoverflow.com/story/sixinghuang> (Top 20% python answerer)

Pluralsight IQ

<https://stackoverflow.com/users/story/266852>

Python: Proficient
Applied Data Mining with Python: Expert

Blog

<http://dgg32.blogspot.com/>

Project Euler

Level 4

DATA SCIENCE PROJECTS

Project

DNA vs RNA simulation: to reveal improper data preparation can lead to artifacts in published scientific journal articles. Using different probability distributions to validate the robustness of the conclusion.

Kaggle: Predict future sales: Using Xgboost to predict the next monthly sales in Russian software firms - 1C Company. It is also the graduation project of a Coursera course. Submission ranking: 2981 out of 8791 teams.

Studies on Digital Sequence Information on Genetic Resources: Systematic and statistical analysis of the digital sequence information on public databases. Reviewed by governments and stakeholders.

Kaggle: IEEE-CIS Fraud Detection: Using Lightgbm to predict credit card fraud based on data provided by Vesta Corporation. Submission ranking: 2849 out of 6381 teams.

BacDive: automatic tagging of habitat description: Combining techniques such as TF-IDF, Keras' LSTM and word embedding to predict the three-levels tags from new unseen habitat text descriptions in scientific literature.

CDiff: Data exploration of the CDiff genome metadata and use Xgboost to predict whether or not a genome may contain antibiotic resistance gene

Link

<https://www.kaggle.com/dgg321982/rna-dna-different-underlying-distribution>

[https://github.com/dgg32/machine-learning/blob/master/inal_submission.ipynb](https://github.com/dgg32/machine-learning/blob/master/final_submission.ipynb)

<https://www.cbd.int/abs/DSI-peer/Study-Traceability-databases.pdf>

<https://www.kaggle.com/dgg321982/lightgbm-baseline-small-file?scriptVersionId=43211348>

https://colab.research.google.com/drive/1YymsBBvt-b5jEWIC15ZfJmj5d_NrQUX

<https://colab.research.google.com/drive/1Iqb8Cq1iIRxbuxHxHBHqHlcMWmUwxNvH>

