

DNA Barcoding and Phylogenetic Analysis of Plant Species: Urban Barcode Project of New York City

Irving Estevez¹, Stephen Mensah², Rujin Tian³, Martin Fein⁴

¹Biology Undergraduate, LSAMP Scholar, City College of New York, NY, 10031

²Biology Undergraduate, Bronx Community College, City University of New York, NY, 10453

³PhD, Associate Professor, Department of Biological Science, Bronx Community College, City University of New York, NY, 10453

⁴PhD, Chairperson Professor, Department of Biological Science, Bronx Community College, City University of New York, 10453

ABSTRACT

DNA barcoding relies on the use of a standardized DNA region as a tag for simple, rapid and affordable species identification. To get hands-on experience on species identification using molecular tools and to explore the genetic biodiversity of New York City, we participated in the Urban Barcoding Project conducted by the DNA Learning Center of Cold Spring Harbor Lab (CSHL). The gene region that is proposed as the standard barcode for plants by CSHL is a ~600 base pair fragment from the RuBisCo (ribulose biphosphate carboxylase/oxygenase) large subunit (*rbcL*) located in the chloroplast. We began our investigation by collecting and selecting a total of 10 native plant specimens located in the campus of the Bronx Community College of New York City with the help of two mobile apps (Google Maps and Garden Compass). Next, we successfully optimized the protocols provided by the CSHL to achieve DNA purification, *rbcL* amplification and sequencing. Finally, we applied bioinformatic tools (sequence alignment; substitution rate and time computation; 3D structure comparison) for DNA-based species identification, protein structure homology modeling and phylogenetic analysis. Our research experience helped us develop a greater appreciation for the DNA sequence based modern taxonomy in urban environments while gaining an introduction to bioinformatics tools.

Keywords: DNA Barcoding, Garden Compass App, Phylogeny, Protein Superimposition, *rbcL*, Sequence Alignment, Species Identification

I. INTRODUCTION

We are living in an ever increasingly urbanized world, where greater than half of the world's population lives in urban areas. Urban settings have different selective pressures from those on wild habitats. Habitat alteration ranging from land-use transformation to changes in vegetation structure is believed to be responsible for the extirpation of many native plants and animals from urban settings. Investigating urban impact on species is a difficult, complex process. Species lists and data are often limited, especially for invertebrates, fungi, and non-vascular plants that constitute the vast majority of the unknown species in the US. A major challenge to the

mapping of species is the difficulty in identifying them.

DNA barcoding is a molecular approach to identifying species by a short DNA sequence from a uniform location on the genome, rather like a supermarket scanner reads the barcode of an item. It was first proposed and developed in 2003 by a Canadian biologist, Dr. Paul Hebert. In 2004 the Consortium for the Barcode of Life (CBOL) was created as an international initiative dedicated to supporting the development of DNA barcoding as a global standard for species identification. Since then more than 200 organizations from more than 50 countries have joined CBOL and agreed to put their barcode data in a public database known as Barcode of Life Data Systems (BOLD, <http://www.boldsystems>).

org). To raise awareness of DNA barcoding and explore genetic biodiversity in New York City, we participated in the Urban Barcode Project conducted by the DNA Learning Center of Cold Spring Harbor Lab (<http://www.urbanbarcodeproject.org/ubresearch.html>), outreach program for the CBOL.

A region of the chloroplast gene *rbcL* – RuBisCo (ribulose biphosphate carboxylase/oxygenase) large subunit– is proposed by the CSHL for plant DNA barcoding. This fragment has been proved to be sufficiently variable to discriminate among most land plant species. Here we report the application of the DNA barcoding method to identify the native plant samples across our campus. We further explore patterns of genetic variation from taxonomic, phylogenetic, and structural perspectives.

II. METHODS AND MATERIAL

Specimen Collection/Location

We used a smart phone (Apple iPhone 6) to capture a photo of the plant and its environment. We also used a geolocation application (Google Maps) to determine its position on campus. Those two features allow us to form an idea of the distance between the different samples and also to refer back to it in the future. We separated all unknown samples into their own plastic zip locks to maintain its vitality. All samples were stored for less than 2 hours before DNA isolation begins.

Pre-Sequencing Identification

We used “Garden Compass” mobile application as a tool to determine the species name and background. The application uses a plant recognition software with a large plant characteristic data base to match the plant in subject with the closest superficial similarities. You will choose from a list of options. A second option within the application allows you to send the photo to an expert in case no match has been found, in which he/she will respond within 24 hours.

A. DNA Extraction

10 unknown plant leaf samples were used. Each individual specimen was macerated using sterile 1.5 mL microcentrifuge and plastic pestle. Standard silica

protocol from Cold Spring Harbor (http://www.dnabarcoding101.org/protocol_isolating_dna.html#standard) was used for DNA isolation. The concentration and quality of the DNA sample was determined with gel electrophoresis and NanoDrop 1000 spectrophotometer (Fisher Thermo Scientific).

B. PCR Amplification and Sequencing

A total volume of 20µL PCR master mixture contained the following: 2µL 10X PCR buffer, 2µL 2.5mM dioxynucleoside triphosphate (dNTP), 5µL 10µM primer provided by Cold Spring Harbor, 8µL of genomic DNA template, 0.2µL Taq polymerase with 2.8µL of distilled water. The primer pairs *rbcLF* (5’_TGTAACGACGGCCAGTATGTCACCAACAACA GAG ACT AAA GC_3’) and *rbcLR* (5’_CAGGAAACAGCTATGACGTAAAATCAAGTC CACCRCG_3’) were used for the PCR. The PCR was performed with a 2720 thermal cycler (Applied Biosystem) as follows: 94°C for 2 minutes, followed by 35 cycles of 94°C for 30 seconds, 54°C for 30 seconds and 72°C for 1 minute, followed by an elongation step at 72°C for 5 minutes. 1% agarose gel using 1X TAE buffer containing 0.5µg/ mL EB (Ethidium Bromide) was used for PCR product electrophoresis. Gel images with UV camera were obtained using BIO–RAD Gel Doc XR+. The PCR product sizes were determined using 20µL of Gene Ruler 1kb Plus DNA Ladder. We use GENEWIZ DNA/RNA reading services to provide back the genetic sequences.

Post-sequencing Software and Tools Used

A. BLAST

For comparative analysis we used the NCBI website to perform a BLAST search with the sequences provided from GENEWIZ. This BLAST will provide closest, if not identical *rbcL* sequences stored in the GenBank. We considered a successful match if there is a query identity score of >95% involved a single genus.

B. BOLD Systems

The Barcode of Life Data Systems is designed to support the generation and application of DNA barcode data. It accepts sequences *rbcL* and Maturase K genes

(matK) and returns a species-level identification when possible

(http://www.boldsystems.org/index.php/IDS_OpenIdEngine).

C. Phylogenetic Tree

<http://phylogeny.lirmm.fr/phylo.cgi/index.cgi> is a free, simple to use web service dedicated to reconstructing and analyzing phylogenetic relationships between molecular sequences. We used a phylogenetic tree model to get an idea about the rate of mutation the *rbcL* gene undergoes relative to number of years.

D. Protein Data Bank (Pdb) and Uniprot

The Protein Data Bank (<http://www.rcsb.org/pdb/home/home.do>) is a large database that contains 3D structural data of large biological molecules, such as proteins and nucleic acids. We used it to gather 3D structures of RbcL or closest homologues structures if the original has yet to be solved.


E. Chimera

UCSF Chimera is a highly extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles

(<http://www.cgl.ucsf.edu/chimera/docs/ContributedSoftware/matchmaker/matchmaker.html>).

III. RESULTS AND DISCUSSION

We used two mobile apps, Google Maps and Garden Compass, to help streamline the process of plant sample selection, record keeping and preliminary identification (Fig. 1). We selected a total of 10 native plant samples of different species across our campus considering DNA barcoding can only be used if the genetic variation between species exceeds that within species.

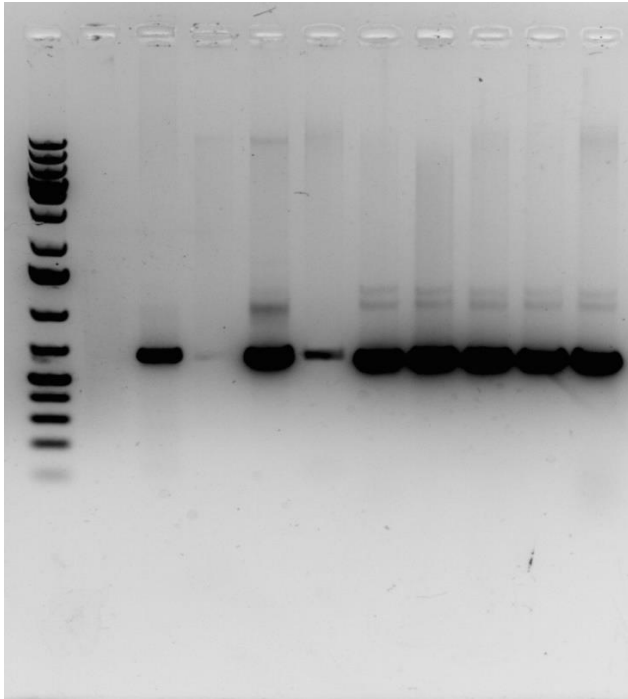


Question: Please identify this plant. →

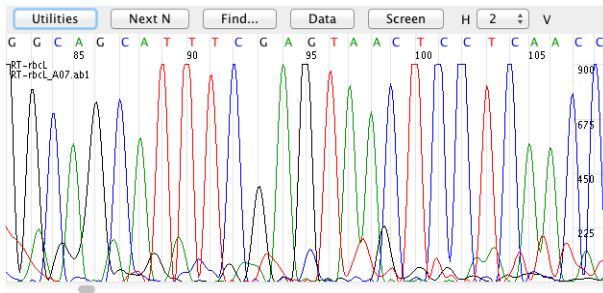
Garden Compass Answer: Your plant appears to be common dandelion (*Taraxacum officinale*), a widely distributed perennial herbaceous plant found in lawns and other open grassy places, fields and road sides. They are deeply rooted plants of about 5-45 cm or 2-18 inches tall. Its name originates from the French word “Dent-de-lion”, meaning ‘the lion tooth’. Its leaves are about 3-12 inches long and with 3 a half inch width, always growing in a basal rosette. Their flower head consists of tiny ray petals. Dandelion is an edible plant; its leaves are consumed as part of vegetable salads. Common dandelion is used for medicinal purposes as well.

Figure 1: Plant Sampling and Selection. Specimens selected for DNA extraction were located and recorded by Google Maps. “Garden Compass” is a free app based on traditional taxonomy and designed to help identify or verify unknown plant species. By simply taking a picture of the plant, the app will pair it with the closest match within 24 hours. Above is a typical feedback received from this app.

We then followed the CSHL protocol for plant DNA extraction from fresh leaves and *rbcL* amplification. The amplified *rbcL* genes of our samples were subsequently sequenced (Fig. 2) for further identification (Fig. 3).



Trace File: RT-rbcL.ab1



Sequence File: RT-rbcL.seq

```
>RT-rbcL_A07.ab1
NNNNNNNNNNNTTNNNTTCACTGGTGAAGATTATNNNNNGACTTATTACTCCTGACTATGAAACCAAGGACTGATATTT
TGGCAGCATTTTCAGTAACCTCAACCTGGAGTCCGCTGAAGAAGCAGGGCCGAGTAGCTGCGGAATCTTCTACT
GGTACATGGACAACGTGTGGACCGATGACTTACGAGCCTTGATCGTTACAAAGGCGATGCTATGGAATTGAGCCTGT
TCCTGGAGAAGAAAGTCAATTTATTGCTTATGTAGCTTACCCATTAGACTTTTTGAAGAAGTTCTGTACTAACAGT
TTACTTCCATTGTAGGTAATGATTTGGTTCAAAGCCTGCGTCTACGTCTGGAAGAATTCGGAATCCCTGTTGG
TATGTTAAAACCTTCAAAGTCCGCTCACGGCATCAAAGTTGAGAGAGATAAATTGAAACAAGTATGTCGTCTCTGTT
GGGATGACTATTAAACCTAAATTTGGGTTTCCGCTAAAAACACGGTAGAGCTGTTATGAATGCTTCCGCGGTGGAC
TTGATTTTACGTCATACNGGTTTCTGANNCGTCNTTTTTCAANNNGGANNNNNTTCAANNNGGANNNNNNNNNNNN
NGGANNCCNNGANNNTTCTGCTNNNNNNNNGANNNNNGANNNT
```

Figure 2. DNA from our 10 plant samples were successfully PCR-amplified and sequenced. The barcode region of rbcL is close to the 625bp band of our 1kb DNA marker.

Taraxacum alpinum ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit gene, complete cds; chloroplast
 Sequence ID: gb|KF602095.1| Length: 1440 Number of Matches: 1

Range 1: 56 to 599 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
994 bits(538)	0.0	542/544(99%)	0/544(0%)	Plus/Plus
Query 1	ATTATAAATTGACTTATATACTCCTGACTATGAAACCAAGGACTGATATTTGGCAG	60		
Sbjct 56	ATTATAAATTGACTTATATACTCCTGACTATGAAACCAAGGACTGATATTTGGCAG	115		
Query 61	CATTTCGAGTAACCTCAACCTGGAGTCCCGCTGAAGAAGCAGGGCCGAGTAGCTG	120		
Sbjct 116	CATTTCGAGTAACCTCAACCTGGAGTCCCGCTGAAGAAGCAGGGCCGAGTAGCTG	175		
Query 121	CCGAATCTCTACTGGTACATGGACAACCTGTGTGGACCGATGACTTACGAGCCTGATC	180		
Sbjct 176	CCGAATCTCTACTGGTACATGGACAACCTGTGTGGACCGATGACTTACGAGCCTGATC	235		
Query 181	GTTACAAGGGCGAGTCTATGGAATTGAGCCTGCTCCCTGGAGAAGAAAGTCAATTTATTG	240		
Sbjct 236	GTTACAAGGGCGAGTCTATGGAATTGAGCCTGCTCCCTGGAGAAGAAAGTCAATTTATTG	295		
Query 241	CTTATGTAGCTTACCCATTAGACCTTTTGAAGAAGGTTCTGTTACTAACCACTTTACTT	300		
Sbjct 296	CTTATGTAGCTTACCCATTAGACCTTTTGAAGAAGGTTCTGTTACTAACCACTTTACTT	355		
Query 301	CCATTGTAGTAAATGATTTGGGTTCAAAGCCCTCGCTGCTACGCTCGGAAGATTTCG	360		
Sbjct 356	CCATTGTAGTAAATGATTTGGGTTCAAAGCCCTCGCTGCTACGCTCGGAAGATTTCG	415		
Query 361	GAATCCCTGTCCCTAATGTTAAACTTCCAAAGCCCTCCGCTCACGGGATCAAGTTGAGA	420		
Sbjct 416	GAATCCCTGTCCCTAATGTTAAACTTCCAAAGCCCTCCGCTCACGGGATCAAGTTGAGA	475		
Query 421	GAGATAAATTGAACAAGTATGTCGTCCTGTTGGAGTGTACTTAAACCTAAATGTC	480		
Sbjct 476	GAGATAAATTGAACAAGTATGTCGTCCTGTTGGAGTGTACTTAAACCTAAATGTC	535		
Query 481	GTTATCCCTAAAACCTACGGTAGACCTTTATGAATGCTTCCGCGGACTGATTT	540		
Sbjct 536	GTTATCCCTAAAACCTACGGTAGACCTTTATGAATGCTTCCGCGGACTGATTT	595		
Query 541	TTAC 544			
Sbjct 596	TTAC 599			

Match Rank	Phylum	Class	Order	Family	Genus	Species	Score	Similarity	E-Value	Status
1	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	pratense	552	99.82	0	Published
2	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	pratense	552	99.82	0	Published
3	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	pratense	552	99.82	0	Published
4	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	pratense	552	99.82	0	Published
5	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	pratense	552	99.82	0	Published
6	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	pratense	552	99.82	0	Published
7	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	pratense	552	99.82	0	Published
8	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	pratense	552	99.82	0	Published
9	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	pratense	551	99.82	0	Published
10	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	pratense	551	99.82	0	Published
11	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	incarnatum	549	99.64	0	Published
12	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	scabrum	547	99.46	0	Published
13	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	arvense	544	99.1	0	Published
14	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	ligusticum	543	99.1	0	Published
15	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	squamosum	543	99.1	0	Published
16	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	arvense	543	99.1	0	Published
17	Magnoliophyta	Magnoliopsida	Fabales	Fabaceae	Trifolium	stratum	543	99.1	0	Published

Figure 3: BLAST and BOLD were performed to make the unambiguous identification of the plant species. We considered the species identifiable with the 600 bp rbcL fragment when BOLD returned a above 99% species level match and when the sequence's closest match in BLAST was unambiguously and exclusively one of the CITES listed species.

BLAST and BOLD both verified the identities of experimented species which initially, have been indicated by the Garden Compass App. At this stage we can confidently state that all unknown plant samples were identified to the lowest taxon possible (Unknown 1: Trifolium Incarnatum; Unknown 2: Pottia Intermedia; Unknown 4: Trifolium Pratense; Unknown 5: Festuca Pratensis; Unknown 6: Chrysanthemum Morifolium; Unknown 7: Chrysanthemum Maximum; Unknown 8: Lactuca Sativa; Unknown 9: Taraxacum officinale (alpinum); Unknown 10: Taraxacum erythrospermum; Unknown 11: Festuca Arundinacea).

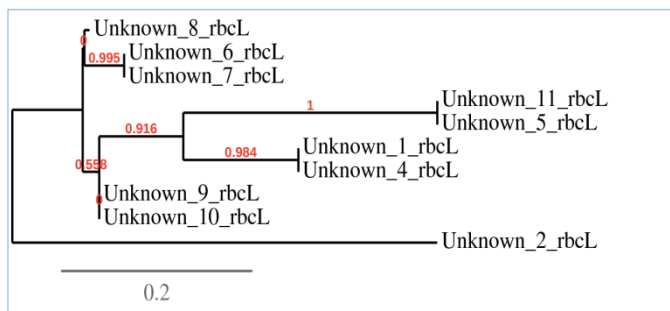


Figure 4: Phylogenetic, or evolutionary relationship of our 10 plant samples based on nucleotide alignment. Horizontal branches represent evolutionary lineages changing over time. The red number is the nucleotide substitutions per site and branch lengths are measured in millions of years.

The partial nucleotide sequence of *rbcL* (~ 600bp) was further used to assess the evolutionary linkage of our samples as presented by the phylogenetic tree (Fig. 4). While identification of some of our samples was more outstanding such as Unknown 8 and Unknown 2, other samples were identified with closer overlaps such as Unknown 6 and Unknown 7. This indicated that for species of close range with this marker, a second marker (e.g. *matK*) might be needed for greater certainty. In addition, the tree is based on substitution rate of a single gene (analyzing local sequence similarities of the *rbcL* gene) to represent the evolutionary distance among all identified species. It tends to overstate both the extent of the inconsistencies and their implications for phylogenetic reconstruction. 3D protein structures, on the other hand, are better conserved through evolution than DNA sequences, though they also change during evolution in response to mutations. Therefore, we performed protein alignment to further investigate the homology divergence of the highly structurally conserved RbcL protein. We predict that between sequences that diverged only recently from their common ancestor, e.g. Unknown 6 (nearest homologue is 1UPP.pdb with ~95.16% identical on the amino acid level) and Unknown 10 (nearest homologue is 4MKV.pdb with ~96.53% identical on the amino acid level), DNA based alignment (Fig. 4) to determine phylogeny is more sensitive since most mutations will be silent (synonymous) and result in little or no change in the amino acid sequence. To test this prediction, we performed amino acid sequence alignment and 3D structure superposition by model building from the known structure of a homologue (Fig. 5 and Fig. 6).

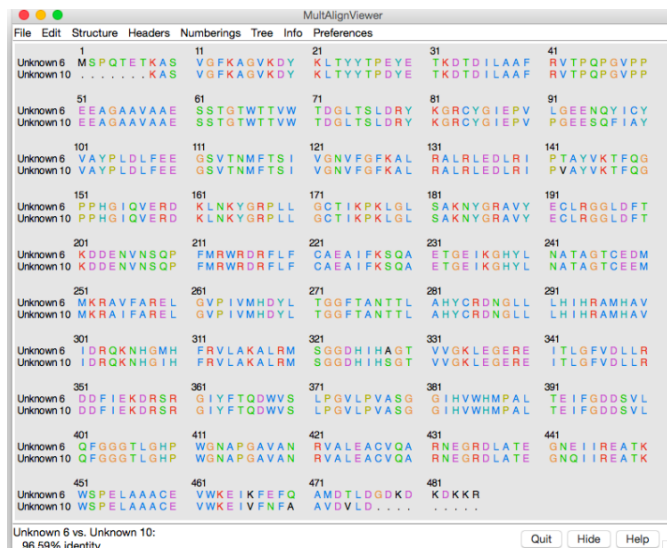
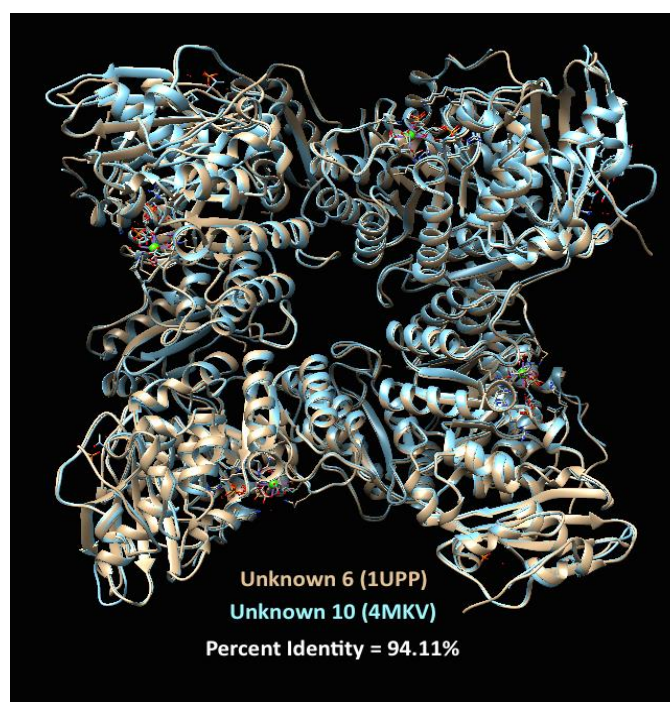


Figure 5: Protein Alignment between Two Samples with Only a Short Evolutionary Distance. The complete amino-acid sequences of RbcL protein of Unknown 6 and Unknown 10 were compared with Chimera Multalign Viewer with strong similarity (96.59%).

Consistent with our predictions, we found a high degree of homology (~97%) between primary structure of proteins encoded by the *rbcL* gene for Unknown 6 and Unknown 10 (Fig. 5). Visualization of 3D superposition with UCSF Chimera was also obtained to better appreciate the homology-derived structural similarity (Fig. 6).



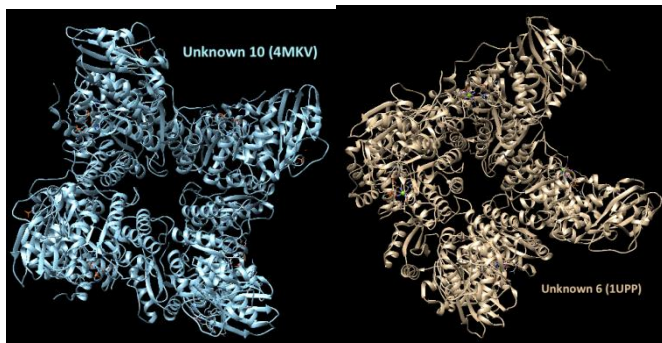


Figure 6: Structural-alignment based superposition. A 3D structure homology model of RbcL protein (octamer, colored as in panel) with Chimera MatchMaker shows strong similarity between Unknown 6 and Unknown 10.

IV. CONCLUSION

Overall, our findings indicated that a single DNA barcode (*rbcL*) offers good, but not outstanding discriminating power for identifying plant species in an urban setting, whereas integrating traditional taxonomy (e.g. Garden Compass) and/or independent genetic markers offer higher resolution. Furthermore, the DNA barcode sequences combined with protein 3D structure analysis can be used to evaluate phylogenetic hypotheses and make homology assessment.

In the future, we plan to expand the number and diversity of our plant samples with multiple barcodes to further investigate the evolutionary pattern of the native plant species in New York City. Combined with structure analysis and reference to existing literature, our study will help identify some of the biological factors (e.g. climate change, invasion by non-native species, mutagens in the environment) that may have led to the evolutionary pattern of our species as well as discovery of unknown and unexpected species.

V. ACKNOWLEDGEMENTS

The authors thank Dr. Kyeng Lee for constructive comments and revisions. We are grateful to Dr. Antonia Florio from DNA Learning Center of Cold Spring Harbor Lab for DNA isolation reagents and primers, Dr. Na Luo from Columbia Medical Center for technical supports. This work was partially supported by CUNY Scholars Research Program from the Office of the Mayor of New York City. Irving Steel is a LSAMP

(Louis Stokes Alliance for Minority Participation, NSF) undergraduate researcher.

VI. REFERENCES

- [1] Dereeper A.*, Guignon V.*, Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* 2008 Jul 1;36
- [2] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004, Mar 19;32(5):1792-7.
- [3] Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 2000, Apr;17(4):540-52.
- [4] Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. New Algorithms and Methods to
- [5] Anisimova M., Gascuel O. Approximate likelihood ratio test for branches: A fast, accurate and powerful alternative. *Syst Biol.* 2006, Aug;55(4):539-52.
- [6] Chevenet F., Brun C., Banuls AL., Jacq B., Chisten R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics.* 2006, Oct 10;7:439.
- [7] Xiang Z. Advances in homology protein structure modeling. *Curr Protein Pept Sci.* 2006 Jun;7(3):217-27.
- [8] Rosenzweig ML. 1995. *Species Diversity in Space and Time.* Cambridge University Press.
- [9] Hebert P.D., Cywinska A., Ball S.L., deWaard J.R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* 270(1512): 313-21.
- [10] Hebert P.D.N, Stoeckle M.Y, Zemplak T.S, Francis C.M. Identification of birds through DNA barcodes. *PLoS Biol.* 2004;2:1657–1663
- [11] Hollingsworth P.M. et al (2009). A DNA barcode for land plants. *Proc Natl Acad Sci USA* 106(31): 12794-7.
- [12] Li M, Wunder J, Bissoli G, Scarponi E, Gazzani S et al. (2008) Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. *Cladistics* 24: 727–745.
- [13] Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins.* 2009; 77:499-508.