The Infinite Discovery Machine

User Manual v0.1

Binomica Labs

Small Thoughtful Science

What is it?

The Infinite Discovery Machine (IDM) is a biological tool for the generation of random proteins from any source of DNA starting material. The heart of the IDM is a plasmid by the same name which acts as an entry point from which random bits of DNA are integrated.

What does it do?

It makes random proteins! The tool takes advantage of three key enzymes: DNAseI, MlyI, and T4 Ligase to chop up DNA, glue the chop back together randomly, and clone it in such that the new scrambled DNA fragments can be expressed in bacteria. The resulting protein, if in frame with the rest of the expression circuitry, will likely be random enough that it has little homology to proteins found in the natural world.

Why make random proteins?

All life is related through the DNA which encodes it. Phylogenetic analysis shows homology between even seemingly disparate organisms and the theory that all life originated from a common ancestor is widely accepted. Any protein isolated from nature shares sequence and/or functional homology with other proteins and whole protein domains are highly conserved. In short, life rarely works with proteins that have not co-evolved with it at least in some capacity. The starting material for novel proteins is still a seed dating back billions of years.

Now, what would happen if you had a means of cheaply and easily generating entirely new proteins that life has never seen before? What would you do? What could be learned? These questions can be explored using this tool!

Since the mechanism has no described target outcome or output aside from an expressible protein, the possibilities for discovery are…infinite. One could focus on exploring protein folding and add to our model by attempting to crystalize these alien forms of peptides. This could help us better predict protein folding and, if automated, could act as a massive dataset for training AI to predict the form and function of natural and de novo proteins. Another example would be to analyze which sequences are most favorable for expression. If you have a 300aa long protein, the possible combinations are roughly 20^{300}…which is an enormous number! Not all these combinations will yield a functional protein, or even be able to be expressed at all.

Despite these filtering attempts, the subset of this expression space has yet to be defined. We only have one example of life to go by so exploring what life elsewhere could be, in an abstract manner, is very VERY exciting. While this could be applied to more directed purposes both academic and industrial, the spirit of this tool is exploration for personal reasons, whatever they may be. How you wish to apply this tool to your research is up to you, but we’d very much like to hear about it so please share!

Is it safe?

Since the core of the IDM is an integration of practically random DNA, there is a non-zero, though astronomically small chance, of producing a sequence which encodes for a protein that could be acutely harmful to human health or the environment. An unknown-unknown varies in its health and safety policy guidelines depending on the country you are operating in and thus is subject to all local laws and regulations.

The bare bones IDM plasmid will express a Met-His-His-His-His-His-His peptide which has no known toxicology nor function aside from having a larger than average affinity to nickel ions, the basis of nickel chromatography via hexahistidine tag purification. This is a common practice in protein purification and some therapeutics still have their 6xHIS tag present at the point-of-care, though this practice has largely been phased out due to more cost-effective tag cleavage mechanisms.

An IDM plasmid that is run through the protocol and is confirmed to have integrated DNA into the MlyI cut site will express whatever sequence is present there as long as the resulting fragment from ATG to TAA is in frame. At this point, the protein expressed has a high likelihood of not having much or any homology to known proteins so it should be handled as an unknown unknown. To reduce the risk of exposure to the novel protein, always wear proper PPE at all times and transform the plasmid into a strain of bacteria with low protein expression and high plasmid expression to further reduce the levels expressed until more analysis is conducted. I recommend the “NEB Turbo” cell line from New England Biolabs since it has very weak and downregulated protein expression but is optimized for plasmid production and extremely rapid colony formation. PCR verify that integration has occurred before culturing any colonies. If you wish to conduct liquid cultures of prospective colonies, do so in as little liquid medium as possible and autoclave any waste made in the process prior to disposal in regular waste streams. Ideally, waste should be treated like regulated medical waste and thrown into a biohazard red bag for incineration, but this may be overkill and not an available option to all researchers. When in doubt, autoclave in a validated autoclave or pressure cooker (test strip mail-in service is cheap) then bleach for 20 minutes in 10% household bleach and then throw into regular waste streams according to local laws.

Research is ongoing to determine the actual likelihood of producing an acutely harmful protein through computer simulations though this dataset may not offer proper risk assessment because, like stated earlier, there are unknown unknowns at play here and the perimeter of possible expressible proteins have yet to be defined – ironic as it may seem, running an IDM protocol and generating a dataset from a collection of many runs might be the only effective way of determining said perimeter.

There is inherent risk in exploring the natural world and the participant or user of the IDM system must do their due diligence to understand these risks. Binomica Labs is not responsible for any injuries or damages resulting from the use of the IDM system as this information is for educational and research purposes only. Be safe and best of luck!

Materials and Methods

Equipment

- Autoclave or validated pressure cooker
- Incubator capable of holding at 37C
- Incubator-Shaker holding at 37C
- PCR Machine
- Pipettes
- Sterile air hood or dead air box
- DNA Agarose Electrophoresis System
- Protein PAGE Electrophoresis System

Consumables and Reagents

- IDM Plasmid
- DnaseI (NEB)
- MlyI Restriction Endonuclease (NEB)
- T4 DNA Ligase (NEB)
- Quick CIP Alkaline Phosphatase (NEB)
- TSS Buffer for cell transformation
- Pipette tips
- Petri dishes or polypropylene container with airtight lid
- Sterile inoculation loops or toothpicks
- LB Media (Miller Mod) (liquid and solid)
- Taq Polymerase Master Mix or components therein
- M13 primers Forward and Reverse:
  5’-GTAAAACGACGGCCAGT-‘3
  5’-CAGGAAACAGCTATGAC-‘3
- Distilled Water
- DNA and Protein Gel reagents and consumables

1. Isolate feedstock DNA from whatever sample you’d like to use. This can be environmental DNA like river water, dirt, etc., a single organism like E. coli or even yourself if you so choose. The source does not matter as much as the quantity of DNA isolated. As much as possible and ideally as clean as possible as the efficiency of downstream processes can be hindered by dirty DNA.

2. Prepare competent E. coli cells in whatever fashion you prefer, I like the TSS method but to each their own.

3. To a clean 0.2mL PCR tube, add 17uL of feedstock DNA >1000ng, 2uL of DNaseI Buffer 10x, 1uL of DNaseI. Vortex vigorously for 5 seconds.

4. Incubate at 37C for 2 hours (changing the incubation time will lead to less or more fragmentation but 2 hours is more than enough for 1 microgram of feedstock).

5. Heat-kill the DnaseI by incubating at 75C for 15 minutes.

6. Take 2uL of this digested DNA and run it on an agarose gel to detect smearing indicating complete digestion. Theoretical yields for DNAseI are fragments of 4bp or less but this varies in real world applications so a low molecular weight smear would suffice. Incubate for longer if smearing is too showing zones of very high molecular weight. Always

run a gel against a ladder to compare. If gel looks good, place DNA sample on ice and proceed to step 7.

7. To a clean PCR tube, add 17uL of digested DNA, 2uL of T4 DNA Ligase Buffer 10x, 1uL of T4 DNA Ligase enzyme. Vortex and incubate at 16C overnight since the starting fragments are very small and efficiency will be dismal. One can vary this process timewise but your mileage may vary. Empirical data required.

8. The next day, to another PCR tube, add 16uL of IDM plasmid backbone, 1uL of Quick CIP Alkaline Phosphatase, 1uL of MlyI enzyme, and 2uL of Cutsmart buffer. Vortex and incubate at 37C for 1 hour. This process opens the IDM backbone at the MlyI site just in front of the ATG start codon and dephosphorylates the 5' end of this open plasmid backbone to reduce chance of self-ligation later. Heat-kill the MlyI enzyme by incubating sample at 65C for 10 minutes or gel purify the plasmid backbone.

9. Ligate the digested feedstock fragments from step 7 along with the IDM backbone from step 8 by adding 17.5uL of each to a single PCR tube followed by 4uL of T4 DNA Ligase Buffer 10x and 1uL of T4 DNA Ligase enzyme. Vortex briefly and incubate at 16C overnight.

10. The following day, heat shock 2 to 10uL of the resulting ligation into a low protein expression E. coli cell line such as NEB Turbo. Allow for a 5 minute recovery on ice and then plate on 37C pre-warmed petri dishes with LB Agar (Miller Mod) with 100ug of Ampicillin. Allow plates to dry and then incubate upside down in a 37C static incubator overnight.

11. The following morning take as many colonies as you’d like or have formed for colony PCR. I use a sterile pipette tube or toothpick to pick and place each colony into a 1.5mL tube containing 200uL of LB-Amp liquid media. Shake the tubes vigorously for 4 hours at 37C.

12. Using M13 primers and an empty IDM plasmid as a negative control, PCR the colonies using your preferred Taq Polymerase master mix and 2uL of bacteria as template DNA.

13. Run an agarose gel of the resulting PCR reactions and make note of any bands that are visibly higher in size than the empty backbone control. Those have inserts and are candidates for expression studies and further analysis. The empty backbone yields a band of 186bp if amplified using M13 primers so using a low molecular weight ladder may be ideal to spot smaller inserts from the background plasmid.

14. Send the positive PCR reaction samples for sanger sequencing via M13. If you use Genewiz, they already have the M13 primers in stock and you can specify to use in house primers which saves on time and money. If you wish you can send unpurified PCR product and they will purify it for you for a fee or you can liquid purify the samples yourself and save even more money.

15. Wait for sanger results to arrive and then analyze the sequenced fragment. Look for any possible frame issues or stop codons between the ATG and the 6xHIS tag region of the sequenced fragment. Run a protein BLAST on any fragments that are in frame and could

potentially be expressing. Determine if any homology is known and what function similar proteins of that kind of structure may have. A comprehensive list of harmful proteins is not directly available, but we ask that you contribute to said list that we may have a better reference to base these comparisons.

16. Grow up the positive colonies that contain in-frame fragments in liquid LB-Amp overnight at 37C shaking and miniprep them the following day.

17. Transform the candidate plasmids into a protein producing strain like BL21(DE3). In a tandem reaction, transform cells of the same cell line with empty IDM backbone as a negative control. Plate cells onto LB-Amp agar and incubate overnight at 37C.

18. The following morning, colony PCR for IDM presence or blindly trust that your transformation worked (bad practice!), set the positive colonies to grow overnight at 37C shaking along with one flask of cells carrying the empty IDM backbone plasmids. Even an empty IDM plasmid will express protein for ampicillin resistance as well as the MHHHHHH peptide so a negative control for this is mandatory.

19. Using a protein purification kit (I like the one from Gene and Cell Technologies), isolate a crude extract from your candidate plasmid flasks and one from your empty backbone control.

20. Process the extract in preparation for a protein gel as per the instructions of your PAGE system’s manual. Visualize the gel against the control and a ladder to determine the novel protein band from the

background proteins. Note its size and band intensity.

21. Now that you know your protein of interest is expressed, the final step here is actually the threshold to an infinite labyrinth of possible experimental analysis and pipelines. From here on out it’s your call as to how, where, and why to proceed with your sample. I’d suggest proposing your protein to a research group at a university that focuses on crystallography or partner up with a structural biologist to learn more about your new protein. One starting point is to run structural prediction using the free software service provided by the University of Chicago’s RaptorX website. It will spit out a 3D model of your protein based on known homology. If you have no homology, it may give faulty results, but this data could be meaningful downstream. If your protein is of 200 amino acids or less you can try QUARK from the I-TASSER suite which focuses on ab-initio sequences (amino to amino interactions rather than structural homology).

22. Lather. Rinse. Repeat. Report back. Enjoy!

For more info, contact us at sebastian at Binomicalabs dot org!