MOLECULAR BIOLOGY FOR THE ENVIRONMENT

A short EC-US Course in Environmental Biotechnology
Under the auspices of the EC-US Task Force in Environmental Biotechnology

Funding Agencies
The V Framework Programme of the European Commission
The U.S. Department of Energy

Venues
Centro Nacional de Biotecnología (CNB), CSIC
Facultad de Ciencias (Biología), Universidad Autónoma de Madrid
Campus de Cantoblanco, Madrid, Spain

Feb. 2-15, 2003

DO-IT-YOURSELF CD
Advanced laboratory and bioinformatics methods in environmental biotechnology

---

The purpose of this CD is to make available the practical exercises that were part of the "**Molecular Biology for the Environment**" course held in Madrid, Feb. 2-15, 2003 to anyone who could not attend the course and who would wish to repeat these exercises independently.

This CD has been designed to allow anyone with the qualifications required for the above mentioned course (late PhD. or early PostDoc in environmental microbiology) and with access to standard microbiology laboratory facilities to execute the five wet laboratory and the two bioinformatic exercises that constituted the practical work of the course.

Each wet laboratory file (entitled *exp1.doc*, *exp2.doc*, *exp3.doc*, *exp4.doc*, *exp5.doc*) consists of a general description of the experiment, a list of required materials and instruments and, a step by step series of instructions.

The file entitled *metarouter.doc* consists of the first of the two bioinformatics exercises. It includes an introduction, a slide presentation and specific exercises on a bioinformatic tool called Metarouter. To access the Metarouter tool, use the following address, username and password:

http://www.almabioinfo.com/MetaRouter/index.html
user: "ecus"
passwd: "ecus_biodeg"
(If you have trouble using the password, please contact ALMA Bioinformatica SL through the web page indicated.)

The file entitled *proteinpredictions.doc* consists of the second bioinformatic exercise. It includes an introduction, a slide presentation and exercises on the prediction of protein sturcture.

Both bioinformatic exercises make use of various websites for which links are provided. Therefore, a fast speed internet connection and a server which can accomodate searches in large databases are required for the successful completion of these exercises.

Good luck!

The Organizers

EXPERIMENT 1

Isolation of mutants defective in seed colonization

Manuel Espinosa (Estación Exp. del Zaidín, CSIC, Granada)

Background

Plant diseases caused by pathogenic bacteria and fungi are major factors in the decrease of agricultural resources. The identification of new targets for antimicrobial drugs, as well as biological control with antagonistic microorganisms appear as interesting future alternatives in the fight against plant pathogens.

Soil bacteria belonging to *Pseudomonas fluorescens* and *Pseudomonas putida* species show metabolic versatility and a variety of characteristics that make them attractive for environmental and agricultural uses. They can colonize the surface of plant roots and the surrounding soil regions (rhizosphere), in a mutualistic association in which the bacteria obtain nutrients from root exudates. In turn, some bacterial strains can promote plant growth and have biocontrol potential against certain pathogens. The efficiency of biocontrol processes depends on the bacterial ability to colonize the roots and surrounding soil regions (the rhizosphere). Bacterial adhesion to seeds (spermosphere) is often the first step in this interaction since agricultural uses of microorganisms often involve coating seeds with bacterial suspensions. This attached population constitutes the initial “inoculum” for the establishment of the biocontrol agent in the rhizosphere. Thus, adhesion to seeds appears to be a key element for agricultural uses of plant beneficial bacteria given that it determines the subsequent colonization of the root system. Furthermore, spermosphere colonization is also important in pathogenesis processes since seeds are the main dispersion vehicle of plant pathogens such as *Pseudomonas syringae*.

Overview

P. putida KT2440 is a soil bacterium that colonizes very efficiently the rhizosphere of different plants of agricultural interest (tomato, bean and corn among others). The aim of this set of experiments is to obtain mutants defective in adhesion to corn seeds and identify the affected genes. Understanding the molecular mechanisms involved in bacterial colonization of the spermosphere can provide new tools to improve the association between beneficial microorganisms and plants, as well as provide new targets to design antimicrobials to fight against plant pathogens. Mutants will be obtained by random transposon mutagenesis with a mini-Tn*5* derivative carrying a kanamycin resistance gene. Mutants deficient in seed adhesion will be selected, their defect will be analyzed and the transposon insertion site will be determined by random PCR and sequencing.

1) Obtaining mutants by random transposon mutagenesis. The strain we will employ is KT2440, a derivative of the soil bacterium *P. putida* mt-2. It will be mutagenized using the suicide plasmid pUT-Km1. Among other characteristics pUT carries *tnp**, a mutant *tnp* gene of IS50R that encodes the transposase needed for transposition of mini-Tn*5* elements. Mini-Tn*5* transposons have proven to be very useful in the insertion mutagenesis of several gram-negative bacteria. These insertions are very stable and distributed randomly within the chromosome. An insertion on the host chromosome that leads to a loss of function will indicate the implication of this gene in the phenotype of interest.

2) Selection of mutants deficient in seed adhesion. The method that we will employ to select mutants that show a reduced seed adhesion capacity will consist in passing the pool of previously obtained Km^R mutants through a seed column. This method has been used before as an enrichment step. However, in this case it will be the primary selection step, allowing sufficient time for those bacteria that retain their colonization capacity to attach to the seeds. Mutants that have lost this capacity will be washed away when the column is opened, and will be recovered by plating in selective medium.

3) Seed attachment assays. To confirm the defects in seed adhesion of the mutants recovered from the columns, three of these mutants will be used, along with the wild type strain, to test their ability to colonize seeds. For that purpose, seeds will be inoculated with a defined amount of bacterial cells. After incubation, the seeds will be washed and the attached bacteria will be recovered by a mechanical treatment, using glass beads. The number of inoculated bacteria and the number of recovered cells will be determined by plating dilutions in selective medium. KT2440R, a KT2440 derivative which is resistant to rifampicin due to a point mutation in the *rpoB* gene will be employed as a positive control for attachment.

4) Identification of the transposon insertion site. The gene interrupted by the transposon insertion will be determined in the mutants showing defects in attachment. The method employed for that purpose will be arbitrarily-primed PCR (or arbitrary PCR), followed by automated sequencing. The method consists in two consecutive PCR reactions. In the first round, DNA (or a colony) of the selected strain is used as a template with an arbitrary primer (ARB1A: 5'-CCACGCGTCGACTAGTACNNNNNNNNNNGATAT-3') and an internal primer of mini-Tn*5*, reading outwards the right end (TNEXT2: 5'-CTTTATTGATTCCATTTTACACT-3'). A second round of amplification is done using 5 µl of the first-round reaction as template. Primers used for the second round are those corresponding to the conserved region of ARB1 (ARB2A: 5'-CCACGCGTCGACTAGTAC-3'), and a second internal primer of mini-Tn5, closer to the end (TNINT2: 5'-CCTGCAGGCATGCAAGCTTCGGC-3'). The PCR products are electrophoresed and the most intense bands isolated with the Qiagen gel extraction kit, and sequenced, using oligonucleotide TNINT as a primer. The ~40 bp distance between this primer and the end of the mini-Tn*5* provides an internal control to ensure that the sequence obtained corresponds to the junction between the transposon and the chromosome.

Experimental procedures

Strains: *P. putida* KT2440 and KT2440R
E.coli CC118λ*pir* (pUT-Km1)
E.coli HB101(RK600)

Materials: Corn Seeds
LB medium
M9 minimal medium for *Pseudomonas*.
Ampicillin, kanamycin and chloramphenicol stock solutions.
0.5 M Benzoate stock solution, pH 7
LB, LB-Km, and M9-benzoate-Km agar plates
Eppendorf tubes, PCR tubes, tips, micropipettes, parafilm, tweezers.
Glass beads (Ø 3 mm), sterile toothpicks. Glass culture tubes.
Incubators at 30ºC, 37ºC and 45ºC.
Microcentrifuge. Vortex.
Thermocycler.
PCR buffer, Taq polymerase, dNTPs, oligo ARB1, oligo TNEXT, oligo ARB2, oligo TNINT.
Electrophoresis equipment, agarose, loading buffer, TAE buffer
UV transiluminator, scalpels or razor blades.
DNA extraction kit.

DAY 1 (time of day given as a suggestion only)

11:30 Inoculate strains in LB:

receptor:	KT2440	
donor:	CC118λ*pir* (pUT-Km1)	kanamycin, ampicillin (Km, Ap)
helper:	HB101(RK600)	chloramphenicol (Cm)

Grow *P. putida* at 30ºC and *E.coli* at 37ºC

18:30 Mating:

1. Pipet 700 µl of receptor strain into an eppendorf tube.
2. Incubate for 10 min. at 45ºC.
3. Add 200 µl of donor strain and 100 µl of helper strain. Mix well.
4. Spin for 1 minute. Discard supernatant.
5. Resuspend pellet in 50 µl of fresh LB.
6. Spot the whole volume in the center of an LB plate. Allow the plate to stand on the benchtop for 5 min.
7. Incubate overnight at 30ºC. DO NOT INVERT THE PLATE.
8. Plate dilutions of the receptor strain on LB plates.

DAY 2 (time of day given as a suggestion only)

11:30 Selection of Kan^R mutants:

1. Pour 1ml of LB on top of the spot of cells in the mating plate.
2. Scrape off cells with the same blue tip.

3 Tilt the plate and aspire/expel liquid a few times to obtain an homogeneous cell suspension.

4 Collect cell suspension in an eppendorf tube.

5 Plate dilutions (10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}) on M9+benzoate+Km.

6 Incubate overnight at 30ºC.

7 Count the colonies in the plates of the receptor strain and calculate the number of receptor cells used in the conjugation. This number will be used to estimate the transposition frequency.

DAY 3 (time of day given as a suggestion only)

9:00 Recovery of Km^R mutants

1 Count colonies and estimate the transposition frequency.

2 Add 2 ml of M9 to the plate containing the highest number of individual colonies. Using a glass spreader, scrape off colonies.

3 Collect cell suspension in an eppendorf tube. Mix suspensions from all groups in one glass tube. Vortex.

4 Inoculate 10 µl in a tube with 2 ml of M9+benzoate+Km. Incubate at 30º until the afternoon.

17:30 Selection of seed adhesion mutants

1 Prepare a seed column by filling a 20 ml syringe with hydrated seeds. Close bottom with Parafilm.

2 Prepare serial dilutions of the mutant mix down to 10^{-5}.

3 Incorporate 1 ml of the dilute mutant mix to the first column. Close the top of the column with Parafilm.

4 Incubate column 1h at 30ºC.

5 Open bottom of column and place on top of a sterile tube. Open the top and collect the flow-through. Add 1 ml of fresh M9 to column to recover any retained but unattached cells.

6 Plate dilutions in LB+Km (10^{-1}, 10^{-2}, 10^{-3}). Incubate overnight at 30ºC.

DAY 4 (time of day given as a suggestion only)

11:30 Seed attachment – preparation

Pick your favorite 3 colonies and inoculate in 2 ml LB+Km.

Inoculate also KT2440R (a Rif^R derivative of KT2440) in LB+Rif.

Grow at 30ºC until the afternoon.

17:30 Seed attachment assays

1 Place 4 seeds in sterile glass tubes (1 seed / tube).

2 Transfer 5 µl of each culture to 1 ml of M9, vortex and incorporate to the tubes containing the seeds.

3 Incubate at room temperature for 45 min.

4 During the incubation time:

a) Label and prepare 8 glass tubes, each with 1 ml of M9. Add 10 glass beads (Ø 3 mm) to one half of the tubes.

b) Plate serial dilutions of the cultures to calculate the size of the initial inoculum.

c) Prepare 4 eppendorf tubes with 90 µl M9.

5 Open the tubes and recover the seeds by flipping the tube on an empty Petri dish. Pick up the seed with tweezers and put it in the tube without glassbeads.

6 Vortex for 30 seconds and repeat step 5, this time placing the seed in the tube that contains the glassbeads. Repeat this process for each tube.

7 Vortex the tubes with the seeds and glassbeads for 1 minute at top speed. Remove 100 µl and plate on selective medium (LB+Rif or LB+ Km). Remove 10 µl, add to the eppendorf tubes with 90µl M9. Mix well and plate.

8 Incubate plates at 30ºC overnight.

DAY 5 (time of day given as a suggestion only)

12:00 Seed adhesion – results

Count colonies in the plates and estimate the percentage of attached *versus* inoculated cells.

12:30 Arbitrary PCR – 1st round

1 With a sterile toothpick, transfer one colony of the mutant(s) showing reduced attachment to a PCR tube containing 20 µl H_2O. Be sure to avoid carrying traces of agar, since it may inhibit amplification.

2 Prepare PCR reactions as follows (per tube):

dNTPs (1mM)	10µl
10x PCR buffer	5µl
oligo ARB1	1µ
oligo TNEXT	2µl
Taq polymerase	2 U > *add last of all*

Complete to a final vol. of 50 µl with H_2O

3 Perform PCR reaction as follows:

95ºC	5 min.	
95ºC	30 sec.]	
30ºC	30 sec.]	x 6 cycles
72ºC	1 min.]	
95ºC	30 sec.]	
55ºC	30 sec.]	x 30 cycles
72ºC	1 min.]	
72ºC	7 min.	
15ºC	∞	

17:30 Arbitrary PCR – 2nd round

1 Transfer 5 µl of the 1st round product to a new PCR tube.

2 Prepare PCR reactions as follows (per tube):

dNTPs (1mM)	10µl
10x PCR buffer	5µl
oligo ARB2	1µl
oligo TNINT	1µl
Taq polymerase	2 U > *add last of all*

Complete to a final vol. of 50 µl with H_2O

3 Perform PCR reaction as follows:

95ºC	5 min.	
95ºC	30 sec.]	
65ºC	30 sec.]	x 25 cycles
72ºC	1 min.]	
72ºC	7 min.	
15ºC	∞	

4 Keep reaction products at –20ºC until electrophoresis

DAY 8 (time of day given as a suggestion only)

11:30 Electrophoresis & isolation of bands for sequencing

1 Cast gel, use combs with large teeth in order to load as much sample as possible.
2 Thaw samples and add 5 µl of loading buffer.
3 Load gel and run electrophoresis at a constant voltage until the front has migrated 50% of the gel length.
4 Stain with ethidium bromide and view under UV.
5 Using a scalpel, cut the band(s) showing highest intensity (generally between 200-700 bp).
6 Put the agarose piece in an eppendorf tube and isolate the DNA with an extraction kit (follow manufacturer's instructions). Resuspend the DNA in the smallest possible volume.
7 Prepare samples to send to the sequencing facility.

DAY 11

Sequence results and bioinformatics

EXPERIMENT 2

Isolation of benzoate degradation mutants of *Pseudomonas putida* mt-2 for the characterisation of the ortho pathway

Silvia Marqués (Estación Exp. del Zaidín, CSIC, Granada)

Background

The aerobic degradation of aromatic compounds proceeds through the formation of a central intermediate compound, the dihydroxylated ring cathecol. The further opening of the ring can take place through two different mechanisms: cleavage between the two hydroxyl groups (*ortho*-cleavage) or cleavage at a position adjacent to one of the hydroxyl groups (*meta*-cleavage) (Figure 1). Bacterial strains have developed degradation pathways that utilize either of the two mechanisms. However, in some strains, several pathways are present simultaneously. This is the case of *P. putida* mt-2, which bears a catabolic plasmid, the so-called TOL plasmid. In this strain, an *ortho*-cleavage pathway is located in the chromosome (*ben* pathway), while a meta-cleavage pathway is coded by the plasmid (*xyl* pathway). In both cases, the pathway expression is triggered by the presence of the substrate, through the activation of a specific regulatory protein, BenR for the *ortho* pathway and XylS for the TOL *meta*-cleavage pathway. The main difference between the two pathways is the range of substrates they can transform. This is due in part to the substrate specificity of the enzymes involved, but also to the limited substrate recognition capacity of each regulator.

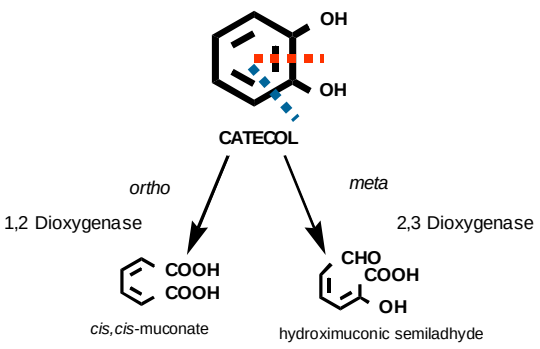


Figure 1: Ortho and meta cleavage of cathecol

The finding and characterization of mutants unable to metabolise the aromatic of interest is a classical approach to analyse a degradation pathway and its regulation. Automation in current laboratory techniques allows a fast and easy application of this approach to obtain, characterise, identify and sequence the genes involved in the degradation of an aromatic compound. With these data, the steps involved in the degradation should be predictable.

Overview

As an example, we will make a random mutagenesis of *P. putida* KT2440 to search for all possible mutants impaired in benzoate degradation, i.e., mutated in the chromosomally encoded pathway. *P. putida* KT2440 is *P. putida* mt-2 strain free of plasmid pWW0. The screening for positive mutants will be carried out automatically in order to analyse more than 5,000 mutants per day. Selected mutations will be directly sequenced to allow a fast identification of the mutated gene.

Random mutagenesis using miniTn5 transposons. MiniTn5 transposons have proven to be very useful in the insertion mutagenesis of several Gram-negative bacteria. *P. putida* KT2440 will be mutagenized with suicide plasmid pUT-Km1. Among other characteristics, pUT carries *tnp**, a mutant *tnp* gene of $IS50_R$ and that encodes the transposase needed for transposition of the mini-Tn5 elements. These insertions are proven very stable and are distributed randomly within the chromosome. An insertion on the host chromosome that leads to dysfunction will indicate the implication of this gene in the phenotype of interest.

1) <u>*P. putida* KT2440 mutagenesis:</u>

a) Day 1: Recipient (*P. putida* KT2440), donor (*E. coli* CC118λ*pir* harbouring plasmid pUT-Km) and helper (*Escherichia coli* HB101 carrying plasmid RK600) strains will be inoculated in LB medium with the corresponding antibiotics. *Estimated time, 30 minutes*.

b) After growth to saturation at 30ºC, prepare a triparental mating of the *P. putida* KT2440 strain with HB101 (RK600) and CC118λ*pir* (pUT-Km). *Estimated time, 1h*.

c) Day 2: Plating for the selection of the Km^R transconjugants. *Estimated time, 2hs.*

2) Selection of mutant strains unable to grow with benzoate as a carbon source: *P. putida* KT2440 bears in the chromosome an ortho-cleavage pathway for the degradation of benzoate (*ben* pathway). The pathway is induced in the presence of the substrate benzoate, through the activation of the regulator BenR. A mini-Tn5 insertion in any of the pathway genes or in any of the regulatory elements would lead to a clone impaired in its ability to grow on benzoate as carbon source.

The screening for the selected phenotype will be carried out automatically using the colony picker robot Q PixII Genetix.

a) Day 3: The Km-resistant colonies (5,000 or more) will be picked in M9 plates with benzoate as a carbon source and in LB-Km plates, and incubated o/n at 30ºC (*Estimated time 4hs*).

b) Day 4: The phenotype of the possible mutants of interest (growth on LB-Km, no growth on benzoate) will be checked again through replica in the same media. (*Estimated time 1h*).

3) Identification of the insertion point in each mutant: We will use arbitrary PCR to identify the insertion point of the transposon as explained for seed adhesion mutants.

a) Day 5: Starting with the colonies of the selected mutants, we will perform arbitrary PCR experiments to identify the insertion point of the Km^R cassette. (*Estimated time 1h to prepare the first PCR, 4 hours for the PCR and 1h for the second. Total time 7hs.*)

*This step will be done together with the localization of the insertion point in the mutants obtained in the seed adhesion experiments (see Experiment 1).

Experimental procedures

Strains: *E. coli* CC118λ*pir* (pUT-Km)
E. coli HB101 (RK600)
P. putida KT2440

Materials: LB medium
M9 minimal medium for *Pseudomonas*.
Ampicillin, kanamycin and chloramphenicol stock solutions.
0.5 M Benzoate stock solution, pH 7
LB, LB-Km, and M9-benzoate-Km agar plates
Eppendorf tubes, PCR tubes, tips, micropipettes, parafilm, tweezers
Sterile toothpicks. Sterile filters. Glass culture tubes.
Colony picker robot Q PixII Genetix (not necessary)
Incubators at 30ºC and 37ºC.
Microcentrifuge. Vortex.
Thermocycler.
PCR buffer, Taq polymerase, dNTPs, oligo ARB1, oligo TNEXT, oligo ARB2, oligo TNINT.
Electrophoresis equipment, agarose, loading buffer, TAE buffer
UV transiluminator.
DNA extraction kit.

DAY 1 (time of day given as a suggestion only)

11:30 Inoculate strains:

Grow:

- *Escherichia coli* CC118λ*pir* (pUT-Km) in LB liquid medium (3 ml) plus rifampicin (20 µg/ml); plasmid pUT-Km has the R6K origin of replication and encodes resistance to ampicillin and kanamycin.

- *Escherichia coli* HB101 (RK600) in LB medium (3 ml) plus Cm (90 µg/ml). Plasmid RK600 is used as a helper; it encodes resistance to chloramphenicol and provides the *tra* functions for the mobilization of the pUT plasmid.

- *P. putida* KT2440 in LB liquid medium (3 ml).

We will incubate these strains at 30ºC for 6-8 hours shaken on an orbital platform operating at 200 strokes per minute.

18:30 Triparental mating:

1. Centrifuge 1 ml of each culture (3 minutes, 13.000 rpm). Discard the supernatant.
2. Add 1 ml of minimal medium (M9), resuspend the cells and centrifuge again.
3. Add 1ml of minimal medium (M9), resuspend the cells and centrifuge again.
4. Discard the supernatant.

5 Resuspend the cells with 100 µl of M9

6 Mix the three strains in an eppendorf tube.

7 Put a sterile filter on an LB plate and add the 300 µl of cells.

1 Let the cells dry over 15 minutes, close the plate with the lid and incubate them at 30ºC o/n.

DAY 2 (time of day given as a suggestion only)

11:30 Selection of Km^r mutants:

1 Take the filter from the mating plate and soak it in 1ml of M9. Vortex vigorously to detach the cells.

2 Prepare serial dilutions in M9 and plate dilutions 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} on M9+glucose+Km plates to select transconjugants. Drop-plate dilutions 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8} on M9 benzoate plates to count recipient cells. Incubate overnight at 30ºC.

DAY 3 (time of day given as a suggestion only)

9:00 Selection of Km^r mutants:

1 Count the colonies obtained in the mutagenesis experiment as well as the *P. putida* KT2440 receptor cells and calculate the mutagenesis frequency. The mutagenesis frequency (F)= number of transconjugants/number of receptors.

2 With the help of the robot, pick all the Km^R colonies first to M9-benzoate plates and then to LB-Km plates.

DAY 4 (time of day given as a suggestion only)

Identify the clones that grow on LB plus Km but not on M9+benzoate. We will check the phenotype again by growing the putative mutants on LB plus kanamycin (3 ml) and M9 plus benzoate (3 ml) in glass tubes. Incubate o/n at 30ºC with shaking.

DAY 5 (time of day given as a suggestion only)

12:30 Identification of the insertion point of the Km^r cassette.

This will be done together with the mutants obtained in the seed adhesion experiments, following the same protocol (see Friday, 7, Experiment 1).

1. Perform arbitrary PCR to determine the insertion point of the Km^R cassette.
2. We will used the product of the PCR described above as template to perform a second amplification round.
3. Electrophoresis and isolation of bands for sequencing. Using a scalpel or razor blade, cut the band(s) showing higher intensity. Generally they will be between 200-700 bp.
4. Put the agarose fragment in an eppendorf tube and proceed to extract the DNA with an extraction kit (follow manufacturer's instructions). Resuspend the DNA in the smallest possible volume.
5. Prepare samples to send to the sequencing facility.

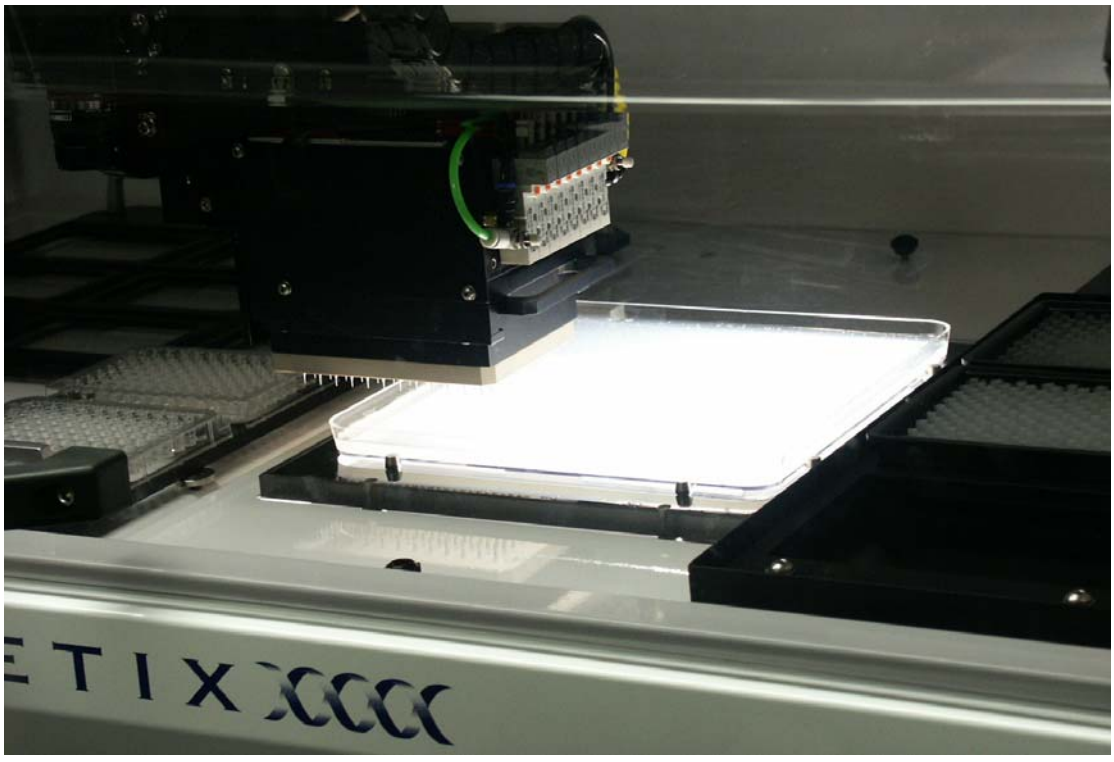


photo by Dan Bost

Robotic colony picker

EXPERIMENT 3

Pseudomonas putida wild type and mutant survival in response to toluene shocks

Ana Segura (Estación Exp. del Zaidín, CSIC, Granada)

Background

Solvent tolerance in bacteria is extraordinary useful in biotechnological applications such as the biotransformation of toxic and water-insoluble compounds in chemicals of added value and, in the removal of pollutants from sites with a large amount of toxic solvents. The knowledge of the molecular bases of solvent tolerance will contribute to a better understanding and better industrial use of these microorganisms.

Although organic solvents with a $logP_{ow}$ value (the logarithm of the partition coefficient of the target compound in a mixture of octanol and water) between 1.5 and 3 are extremely toxic to microorganisms, several *Pseudomonas putida* strains able to grow in culture medium with toluene or related aromatic hydrocarbons have been isolated. *Pseudomonas putida* DOT-T1E is one of these highly solvent tolerant strains. It is widely accepted that organic solvents affect the cell membranes of microorganisms causing an increase in their fluidity. To overcome the damage caused by these solvents, bacteria possess several adaptative mechanisms that readjust membrane fluidity. Among them the *cis-trans* isomerization and the increase in cardiolipin content as a phospholipid headgroup contribute to the rigidification of the cell membrane. Several energy-dependent efflux pumps have been described as key elements in solvent tolerance. Although a lot of work has been done during the last few years, the molecular basis for solvent tolerance is not fully understood. The isolation of mutants of *P. putida* DOT-T1E unable to withstand 0.3% (vol/vol) of toluene and the characterization of the mutation will contribute to a better understanding of this process.

Overview

Survival in response to toluene shocks

Solvent tolerance in *P. putida* DOT-T1E is an inducible process. When a DOT-T1E culture is shocked with 0.3% (vol/vol) of toluene, only one out of ten thousand cells survive the toxic effect of the toluene. However, if the culture has been pregrown in the presence of small amounts of toluene (i.e. toluene in the gas phase), the addition of 0.3% toluene does not decrease the viability of the culture. This means that at least some of the genes involved in solvent tolerance are induced by toluene. We will study the phenotype of several mutants in response to a toluene shock under induced and uninduced conditions.

Day 1: We will grow the wild-type and the different mutants on LB (50ml) with or without toluene in the gas phase.

Day 2: We will dilute the cultures in fresh medium (LB with or without toluene) and grow them until they reach a turbidity of about 0,8 at 660nm. Cultures will be divided in two halves; one of them will be shocked with 0,3% (vol/vol) toluene, and the other will be kept as control. The number of viable cells will be determined before the toluene addition and, 10, 30 and 60 minutes later.

Experimental procedures

Strains:
Pseudomonas putida DOT-T1E (silvestre), RifR
P. putida DOT-T1E-18 (*ttgB:: ´phoA*-km), KmR
P. putida DOT-T1E-28 (*ttgH::* sm), SmR
P. putida DOT-T1E-62 (*ttgV::* km), KmR
P. putida DOT-T1E-42 (*fliP:: ´phoA*-km), KmR
P. putida DOT-T1E-6 (*ttgD:: KilAB*), TelR

Materials:
LB medium
M9 minimal medium for *Pseudomonas*.
Streptomycin, kanamycin, rifampicin and telurite stock solutions.
LB agar plates
Eppendorf tubes, tips, micropipettes, parafilm.
Toluene.
100 ml flasks, glass cylinders for toluene.
2 ml syringes.
Incubators at 30ºC.
Microcentrifuge. Vortex.

DAY 1 (time of day given as a suggestion only)

18:30 Inoculate strains with or without toluene

Add 50ml of LB (plus the corresponding antibiotic) in a 100 ml flask. Inoculate each strain with or without toluene in the gas phase. In the flask with toluene, introduce the toluene in glass cylinder with a syringe (about 100µl toluene/glass cylinder). The toluene should not be in touch in any moment with the liquid medium. Seal the flask with parafilm. Grow the cultures o/n at 30ºC with agitation.

DAY 2 (time of day given as a suggestion only)

11:30 Reinoculate strains with or without toluene

1 Dilute the o/n culture in fresh LB medium plus/minus toluene in the gas phase.

2 Seal the flask with parafilm and grow them in an orbital shaker at 30ºC until the cultures reach an O.D. of 0.8 at 660 nm.

18:30 Survival in response to toluene shocks

1 Draw lines that divide the LB plates in four quarters.

2 Divide the culture in two halves: one will be kept as control without toluene and to the other half we will add 0.3% toluene (vol/vol). Just before the addition of toluene, take 30μl of the cultures and put the drop in one of the quarters of the LB plate (total of three drops).

3 Dilute10μl of culture in 990μl of M9 and prepare serial dilutions (10^{-4}, 10^{-6}). Drop 30μl of each dilution (three drops of 30μl) in each quarter of the plate. Incubate the plate o/n at 30ºC.

4 Add 0,3% toluene (vol/vol) to one flask and take samples of both flasks (with and without 0.3% toluene) as indicated above at time 10, 30 and 60 minutes after the addition of toluene. Incubate the plate o/n at 30ºC. *Estimated time: 1h 30 minutes.*

EXPERIMENT 4

Induction of degradation pathways by aromatic compounds: cascade induction of the TOL *meta*-cleavage pathway of *Pseudomonas putida*

Silvia Marqués (Estación Exp. del Zaidín, CSIC, Granada)

Background

The TOL plasmid pWW0 of *Pseudomonas putida* encodes the information for the catabolism of benzoate and alkylbenzoates through a *meta*-cleavage pathway, in which the aromatic carboxylic acids are first oxidized to the corresponding catechols, which undergo meta-cleavage fission to yield a derivative of muconic acid semialdehyde, which is further metabolized to Krebs cycle intermediates. The genes encoding the enzymes of the meta-cleavage pathway in pWW0 form an operon, under the control of a single promoter, Pm. The *xylS* gene, encoding the regulator of the meta-cleavage pathway, is located in 3' with respect to the meta operon, and it is independently transcribed (Figure 1). The XylS protein is synthesized constitutively at a low level and becomes active to promote transcription from Pm when a benzoate effector, i.e. 3-methyl benzoate, is added to the culture medium. Alternatively, *xylS* can be overproduced when a second regulatory protein, XylR, is induced by a toluene derivative. In these conditions, XylS can stimulate transcription from Pm in the absence of effectors (Figure 1).

Both the regulatory gene *xylS*, and the Pm promoter fused to a reporter gene are available in different plasmids to allow study of the regulatory process in a heterologous background such as *E. coli*. We will use different constructions to test the hypothesis of a cascade induction of the Pm promoter through overproduction of XylS.

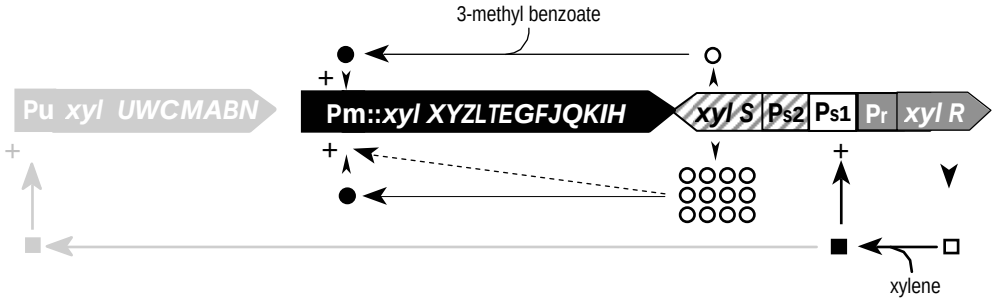


Figure 1: Induction of the TOL meta-cleavage pathway

Overview

E. coli strains bearing a reporter Pm::'lacZ fusion and different sets of regulators, i.e., the *xylS* gene, the *xylR* gene or both *xylS* and *xylR* genes, will be induced by the addition of different aromatic compounds. β-galactosidase activity will be measured and compared.

Day 1: Inoculate cultures. Grow overnight at 30ºC.

Day 2: Culture induction. *Estimated time,4h 30´*.

Sample collection for β-galactosidase activity determination. *Estimated time, 1h*.

Experimental procedures

Strains: *E. coli* MC4100 (pJLR107, pERD103)
E. coli MC4100 (pJLR107, pRL38)
E. coli MC4100 (pJLR107, pRD32)
E. coli MC4100 (pJLR107)

Materials: LB medium
Ampicillin, kanamycin and tetracycline stock solutions.
Aromatic inducers:3-methyl benzoate, 2,6-dimethyl benzoate, toluene.
MATAB solution: 2 mg/ml MATAB in TRIS-HCl 0,2 M pH8.
Z Buffer: 60 mM Na_2HPO_4; 40 mM NaH_2PO_4; 10 mM KCl; 1 mM $MgSO_4$; 50 mM β-mercaptoetanol.
Sodium carbonate 0,5M.
ONPG solution: 4 mg/ml in 0,1 M phosphate buffer, pH7.
Eppendorf tubes, tips, micropipettes, plastic assay tubes.
Glass culture tubes.
Incubator at 30ºC and 4ºC. Water bath at 30ºC.
Timer. Vortex.

DAY 1 (time of day given as a suggestion only)

17:30 Inoculate strains:

Starting from single colonies, inoculate culture tubes containing 3 ml LB/antibiotics with each *E. coli* strain. Grow o/n at 30ºC with shaking.

DAY 2 (time of day given as a suggestion only)

12:00 β-galactosidase induction:

1 Dilute every culture 1/100 in 3 ml of the same medium. Prepare 4 diluted cultures of every strain.

2 Shake at 30ºC for 30 min.

3 Induce by adding 1 mM of one of the aromatic compounds to the culture. Use one culture with no addition as negative control.

4 Shake at 30ºC for 4 hours.

17:30 β-galactosidase determination:

5 Transfer all the cultures to 4ºC.

6 Measure β-galactosidase as follows:

i) In plastic assay tubes, mix 20 µl of each culture with 20 µl MATAB solution.

ii) Incubate 20 min at 4ºC.

iii) Measure A_{660} of 1/10 diluted samples.

iv) Add 800 µl Buffer Z to each tube. Incubate 3 min a 30ºC to equilibrate the temperature.

v) Start β-galactosidase assay by adding 200 µl ONPG substrate solution.

vi) Incubate at 30ºC until a yellow colour develops. Record elapsed time accurately.

vii) Stop the reaction by adding 2 ml sodium bicarbonate 0,5M.

viii) Measure A_{420} and A_{550}.

ix) Calculate Miller Units according to the formula:

$$\text{Miller Unit} = \frac{A_{420} - 1{,}7\ A_{550}}{\text{Time} \times \text{Volume} \times A_{660}} \times 1000$$

EXPERIMENT 5

Bacterial gene expression with DNA microarrays

Víctor Parro (Centro de Astrobiología, CSIC-INTA, Madrid)

Background

DNA microarrays, also known as DNA chips or microchips (Southern et al., 1994; for a review see Nature Genetics 21, supplement, 1999, or see http://industry.ebi.ac.uk/~alan/MicroArray/), are robotized and miniaturized extensions of hybridization-based methods that have been used for decades to identify and quantify nucleic acids in biological samples (Southern and Northern blots, colony hybridizations and dot blots). Samples of interest are labeled and put together with the already known probes on the array. After hybridization and washing, the array is scanned to obtain an image which identifies the presence of individual nucleic acids species in the sample as reflected by the amount of hybridization to the complementary probes immobilized in known position of the array. Hybrids stability is determined by the complementarity degree between the two nucleic acids (probe and target) together with external factors like the ionic strength of the medium, the pH value or the temperature. Using DNA microarrays technology, thousands of molecular probes (cDNAs, PCR fragments, or oligonucleotides) can be covalently tethered to a solid support (glass, nitrocellulose, nylon, etc.) to study differential gene expression or to identify single nucleotide polymorphisms (SNP) in a unique hybridization step. The very recent advances in this techniques have permitted a substantial improvement of the sensitivity so that it is possible to identify minor genomes representing as little as 1% of the total population (Gerry et al., 1999; Favis et al., 2000).

Two basic strategies can be followed for the construction of a DNA chip: i) by directly spotting a probe on a solid support where the probe can be an oligonucleotide, a PCR amplified DNA fragment, a plasmid, a purified DNA fragment or PNAs; ii) by *in situ* synthesis of the probes following the

photochemical deprotection method (photolytography, USA patent number 6.022.963), the ink jet method (Blanchard et al., 1996), or by physically located reactives (Maskos and Southern, 1993). A robotized system, arrayer, is required for directly spotting, and it is possible to print up to 2,500 samples (100 micrometers of diameter each) per cm^2. Using *in situ* synthesis following the photochemical deprotection method, it is possible to obtain more than 65.000 spots (50 micrometers of diameter each) per cm^2.

DNA chips can be applied mainly to gene expression analysis and genotyping. It is possible to analyze gene expression at the mRNA level of thousands of genes in samples extracted from ill tissues (cancer, viral or bacterial infections, etc.), or in the infectious microorganisms (viruses, bacteria, fungi, etc.). New drugs and diagnostic methods can be found and designed after the genes involved in illness development are discovered and their expression patterns deduced. New mutations and single nucleotide polymorphisms (SNP) will be detected and mapped after re-sequencing and genotyping studies (Hacia *et al*., 1998, Kozal *et al*. 1996, Parinov *et al*. 1996; Gerry *et al*. 1999).

Biodiversity is another area of research which can benefit from the use of DNA chips. It is possible to identify different species or even different strains or variants of the same species (Gingeras et al., 1998) for medical (drug resistances, toxins, pathogenic factors, etc.) or ecological purposes (biodiversity, polymorphism distribution, genomic evolution, etc.). Gingeras *et al.* built a DNA-chip with oligonucleotides interrogating all positions in both strands of a 705 bp DNA fragment from *rpo*B gene of *Mycobacterium tuberculosis,* to analyse the existence of possible mutations conferring rifampicin resistance in a set of 63 clinical samples of *M. tuberculosis*. Specific single nucleotide polymorphisms can be identified in a DNA microchip so that each strain or species have its own pattern.

Another example for bacterial identification is the USA patent number 5.925.522, where Wong *et al.* described new methods for the detection of *Salmonella* using specific oligonucleotides DNA chips.

DNA-chips technology is quickly developing, improving the sensitivity and reliability of the detection methods for the discovery of genetic alterations in complex mixtures. Former detection methods were very laborious and expensive. With appropriately designed DNA-chips, it is possible to identify single point mutations in a few hours. With conventional methodology, days or even weeks would be necessary to achieve the same results.

Nanogen Inc. (Gilles *et al.*, 1999) has developed a rapid assay for SNP detection that utilizes electronic circuitry on silicon microchips. This method insures the transport, concentration and attachment of DNA to selected electrodes which act as test or reaction sites, to create an array of DNA samples. By controlling the electric field, it is possible to hybridize fluorescently labelled DNA to the probes on the array and to identify single mismatches. In contrast to nonelectronic or passive hybridization with conventional arrays on paper or glass chips, the use of electronically mediated active hybridization to move and concentrate target DNA molecules accelerates hybridization so that it takes place in minutes rather than hours, as normally required for passive hybridization.

References

Blanchard, A. P., Kaiser, R. J. and Hood, L. E. 1996. Synthetic DNA arrays. Biosensors and Bioelectronics 11, 687-690.

Bulyk, M. L., Gentalen, E. Lockhart, D. J. and Church, M. 1999. DNA-protein interactions by double-stranded DNA arryas. Nature Biotech. 17, 573-577.

Favis R, Day JP, Gerry NP, Phelan C, Narod S, Barany F. (2000). Universal DNA array detection of small insertions and deletions in BRCA1 and BRCA2. Nat Biotechnol. 18(5):561-4.

Gerry ,N.P., Witowski, N.E., Day, J., Hammer, R.P., Barany, G. and Barany, F. 1999. Universal DNA microarry method for multiplex detection of low abundance point mutations. J. Mol. Biol. 292, 251-262.

Gilles PN, Wu DJ, Foster CB, Dillon PJ, Chanock SJ. Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips. Nat Biotechnol. 1999 Apr;17(4):365-70.

Gingeras, T.R., Ghandour, G., Wang, E., Berno, A., Small, P., Drobniewski F., Alland, D., Desmond, E., Holodniy, M., and Drenkow, J. 1998 Simultaneous Genotyping and Species

Identification Using Hybridization Pattern Recognition Analysis of Generic Mycobacterium DNA Arrays. Genome Research. 8:435-448.

Hacia, J.G. 1999. Resequencing and mutational analysis using oligonucleotide microarrays. Nature Genetics 21, 42-47.

Hacia J.G., Sun B., Hunt N., y col (1999), Strategies for mutational analysis of the large multi-exon ATM gene using high-density oligonucleotide arrays, Genome Research 8: 1245-1258.

Kozal MJ, Shah N, Shen N, Yang R, Fucini R, Merigan TC, Richman DD, Morris D, Hubbell E, Chee M, Gingeras TR. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat Med. 1996 Jul;2(7):753-9.

Maskos, U. and Southern E. M. 1993. A novel method for the analysis of multiple sequence variants by hybridisation to to oligonucleotide arrays. Nucleic Acid Res. 21,2267-2268.

Maskos, U. and Southern E. M. 1993. A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesized on a glass support. Nucleic Acid Res. 21, 4663-4669.

Southern *et al.* 1994. Arrays for complementary oligonucleotides for analysing the hybridization behavior of nucleic acids. Nucleic Acids Res. 22, 1368-1373, 1994

Overview

Day 1: RNA labeling and purification of labeled cDNA (3 to 4 h)

Day 2: Pre-hybridization (1h) and hybridization (o/n)

Day 3: Washing (30 min) and scanning (at CAB)

Day 4: Image analysis and discussion of results

1 Manufacture a DNA microarray containing array elements corresponding to genes of interest or from a plasmid library.

2 Prepare a fluorescent cDNA sample from total or purified mRNA by enzymatic labeling.

3 Hybridize the fluorescent sample to the microarray under conditions of high stringency. Labeled probes may lower the hybrid melting point.

4 Wash the microarray to remove unhybridized sample.

5 Scan the microarray to collect fluorescent emission.

6 Quantify the fluorescent emission at each position within the microarray.

7 Assign gene expression values by comparing the experimental data to the appropriate controls.

Experimental procedures

Materials and solutions:

i. Six printed chips (CAB)
ii. DNA free- total RNA samples. Six RNA samples. (CAB)
iii. Cy3-dUTP and Cy5-dUTP for 12 labeling reactions (Amersham)
iv. CyScribe First Strand cDNA Labeling System-dUTP (Amersham)
v. 0,5M NaOH and 0,5N HCl
vi. PCR Purification kit , TE, etc
vii. Bioanalyzer 2100 (CAB)
viii. Speed-vac
ix. Salmon sperm DNA, SDS, BSA, 20XSSC, Hybridization buffer (Telechem), 50 ml falcon tubes.
x. Cover slip slides (GraceBio-Sigma) (CAB)
xi. Six Hybridization chambers (CAB)
xii. Scanner (CAB)
xiii. Analysis software (ScanAlyze). Free download from the web and Adobe-Photoshop software.
xiv. Thermocycler
xv. Water bath
xvi. Microcentrifuge

Print DNA samples onto slides. DONE AHEAD OF TIME

CAB will provide 6 chips with double arrays from a gram- bacteria.

Total purified and DNA free bacterial RNA. DONE AHEAD OF TIME

RNA extraction has been made in the CAB with RNeasy Minikit (Cat. Nº 74104, Qiagen) following instructions of the manufacturer. (In case RNA was obtained with methods based in phenol or other, it would be treated with RNase-free DNase I).

Protocol 1. RNase-free DNase I treatment for RNA samples

---

1 Mix in an eppendorf:

o	Total RNA	30μg	or	100μg
o	5x buffer[a]	20μl		100μl
o	RNase-free DNase I	20-30 U Kunitz (Roche)		30-40 U
o	H_2O up to	100μl (final vol)		500μl

2 Incubate 30' at 37ºC.

3 Mix, centrifuge up and down and change to a new tube.

4 Incubate for further 30' at 37ºC.

5 Phenol extraction (1x PHOH, 1xPHOH/CHCl3, 2x CHCl3 extraction).

6 Ethanol precipitate.

7 Resuspend in DEPC-treated water up to 2-3 □g/□l.

a. 5x buffer: 0.1 M AcNa, 5mM $MgCl_2$, pH 5.5

Quality test for RNA. Bioanalyzer images. DONE AHEAD OF TIME

This test will be made in the Bioanalizer 2100 from CAB.

DAY 1 (time of day given as a suggestion only)

9:00 Label RNA with FluoroLink Cy3d-UTP and FluoroLink Cy5 d-UTP

Protocol 2. Total Bacterial RNA Labeling with Random Hexamers

---

Reagents and Equipment

- Six different samples of RNA purified
- Termocycler
- Speed-Vac
- dNTPs mix (dATP, dGTP, dCTP, dTTP at neutral pH from Invitrogen)
- 0,1M DTT
- CyScribe First Strand cDNA Labeling System-dUTP (Amersham)
- 5x 1st strand buffer
- Autoseq G-50 (Cat. Nº: 27-5340-01, Amersham)
- Water-DEPC (Cat.Nº.9906, Ambion)
- FluoroLink Cy3-dUTP (Cat. Nº. PA 53022 from Amersham) and Fluorolink Cy5-dUTP (Cat. Nº. PA 55022 from Amersham).
- Pipettes 2-10μl, 20-100μl and tips.
- 0,5N HCl and 0,5N NaOH

Method

1 2-25μg of total RNA in 10-14μl of diH_20 should be purified using RNA kit (Qiagen "Rneasy Minikit 50", Cat nº 74104) or hot phenol extraction. After obtaining RNA sample, it is important to get rid of residual DNA with RNase-free DNAse followed by 1x PHOH, 1xPHOH/$CHCl_3$, 2x $CHCl_3$ extraction "RNase-Free DNase set (50) Cat nº 79254 de Quiagen. Extent of RNA purity achieved by PHOH-$CHCl_3$ extraction is absolutely critical for successful labeling. Dry in a speed-vac and dilute in 14μl de water-DEPC.

2 5μg of RNA should be mixed with 500ng of random primers (Invitrogen) on ice in a total volume of 15μl.

3 Incubate at 65 ºC for 10' in a termocycler

4 Incubate on ice for 2'.

5 Add 3μl of 1 mM FluoroLink Cy3 (or Cy5- dUTP)

6 Add 11.6μl and mix everything by pipeting 3-4 times on ice of reverse transcription mix:

   a. 3μl 0,1M DTT
   b. 6μl 5x 1st strand buffer.
   c. 0.6μl from a mix of dNTPs (25 mM dATP, 25 mM dCTP, 25 mM dGTP, 10 mMdTTP).
   d. 2μl reverse transcriptase from CyScribe First Strand cDNA Labeling System-dUTP (Amersham)

7 Mix all components and spin briefly.

8 Incubate the complete mixture and perform the following steps in a termocycler:

   a. 10' at RT
   b. 110' a 42ºC
   c. 10' at 65ºC

9 Stop by adding 3μl of 0,5N NaOH for 10'-30' (for 10-15μg will be 20-25') at 65º C.

10 Neutralize with 3 μl of 0,5M HCl.

11 Add 25μl H2O and purify labeled cDNA products using one of the kits used for the purification of the PCR products (as AutoSeq G-50 from Amersham).

12 Elude DNA with 50μl buffer EB or water.

13 Store the labeled probe at –20^{0}C in a freezer protected from the light.

DAY 2 (time of day given as a suggestion only)

11:30 Quality test for cDNA labeled.

This test will be made in the Bioanalizer 2100 with the RNA lab chip kit from CAB.

Pre-hybridize arrays with a pre-hybridization buffer

Protocol 3. DNA microarray pre-hybridization.

Reagents and Equipment

- Hybridization cassettes (Telechem)
- Cover slip slides (Grace bioLabs) (Cat. Z36,590-4 Sigma)
- Reagents: 20x SSC, BSA 10%, SDS 10%
- Denatured Salmon sperm DNA
- MilliQ water and isopropanol
- Termocycler at 95ºC for denature DNA
- Falcon tubes for 50ml.
- Pipettes 2-10μl, 20-100μl and tips.

Method

1 Prepare prehybridization buffer containing 5xSSC, 0.1% SDS, 0.1 □g/□l of sheared denatured salmon sperm DNA, and 1% BSA (Sigma CatNº A-9418)

2 Put 2-4μl per cm^2 into a cover slide (GraceBiolabs) and then over the slide (20μl) Alternatively, place slides in prehyb. solution or flood the slides with prehyb. sol.

3 Incubate at 42ºC for 30- 45 min.

4 Wash slides by dipping five times in room temperature MilliQ water.

5 Dip the sides in room temperature isopropanol and air dry. (USE SLIDES BEFORE ONE HOUR AFTER DRYING).

12:30 Set hybridization. Hybridize two arrays per slide with a mix of Cy3 and Cy5 labeled probes each one between 4-12 h at 55ºC.

Protocol 4. DNA microarray hybridization.

Reagents and Equipment

- Hybridization cassettes (TeleChem)
- Cover slip slides (Grace Bio-Labs)
- Microarray wash station (TeleChem)
- SDS 10% and buffer 20XSSC
- ScanArray 3000, 4000, or 5000 (GSI Lumonics)
- Pipettes 2-10μl, 20-100μl and tips.

Method

6 Combine purified Cy3 and cy5 labeled probes (6 final vol.) and dried into the speed-vac.

7 Resuspend the dried probes in 1.0 part of aqueous solution (1.5μl of dH_2O + 1.5μl of 10 μg/μl sheared denatured salmon sperm DNA, up to 0.1 μg/μl l final hyb concentration.).

8 Heat the probe mixture at 95 ºC for 3 min to denature. Centrifuge an up and down.

9 Add 4.0 parts (16μl) of UniHyb Hybridization Solution (TeleChem). Pre-warm and mix UniHyb Hybridization Solution for 30 sec at 65 ºC before utilization. Pipette 16μl of fluorescent probe at the middle of a 22 x 22 mm cover slip.

10 Place the microarray slip onto the cover slide to take it by capillarity, such that the sample forms a thin monolayer between the cover slip and the microarray[a].

11 Place the microarray in a hybridization cassette.

12 Add 5.0µl of 5xSSC + 0.2% SDS or water to the slot in the cassette for humidification[b]

13 Seal the hybridization cassette containing the microarray and

14 Submerge the hybridization cassette in a water bath set at 55°C[c].

15 Hybridize for 6 up to 12 hrs at 55°C.

a. Cover slips must be free of oils, dust and other contaminants. Lower the cover slip onto the microarray from left to right so that the sample pushes out air bubbles as it forms a monolayer against the microarray surface. Small air bubbles trapped under the cover slip exit after several minutes at 62°C.

b. Prevents drying of the sample under the cover slip.

c. A temperature of 62°C works well for cDNA-cDNA hybridizations. Lower temperatures should be used for hybridization to oligonucleotides.

DAY 3 (time of day given as a suggestion only)

11:30 DNA microarray washing.

Protocol 5. DNA microarray washing.

---

1. Following hybridization, remove the microarray from the hybridization cassette and place it immediately into the wash station[a]. (50ml falcon tubes)
2. Wash the microarray for 5 min at room temperature in 2X SSC + 0.2% SDS[b] two times.
3. Transfer the wash station and microarray to a second beaker containing 400 ml 0.2X SSC and 0.2% SDS.
4. Wash the microarray for 5 min. at room temperature 0.2 X SSC and 0.1% SDS.
5. Rinse the microarray briefly in a third beaker containing 0.1 X SSC to remove the SDS.
6. Allow the microarrays to air dry.
7. Scan the microarrays for fluorescence emission in a ScanArray 3000.

a. Wash station should be placed in a 600 ml beaker containing 400 ml 1x SSC + 0.1% SDS. The microarray should be transferred quickly from the cassette to the wash station. Leaving the microarray at room temperature will lead to elevated background fluorescence.

b. The cover slip should slid off the microarray during the wash step. If the cover slip does not slid off within 30 sec, use forceps to gently remove it from the microarray surface. Failure to remove the cover slip will prevent efficient washing of the microarray.

NOTES

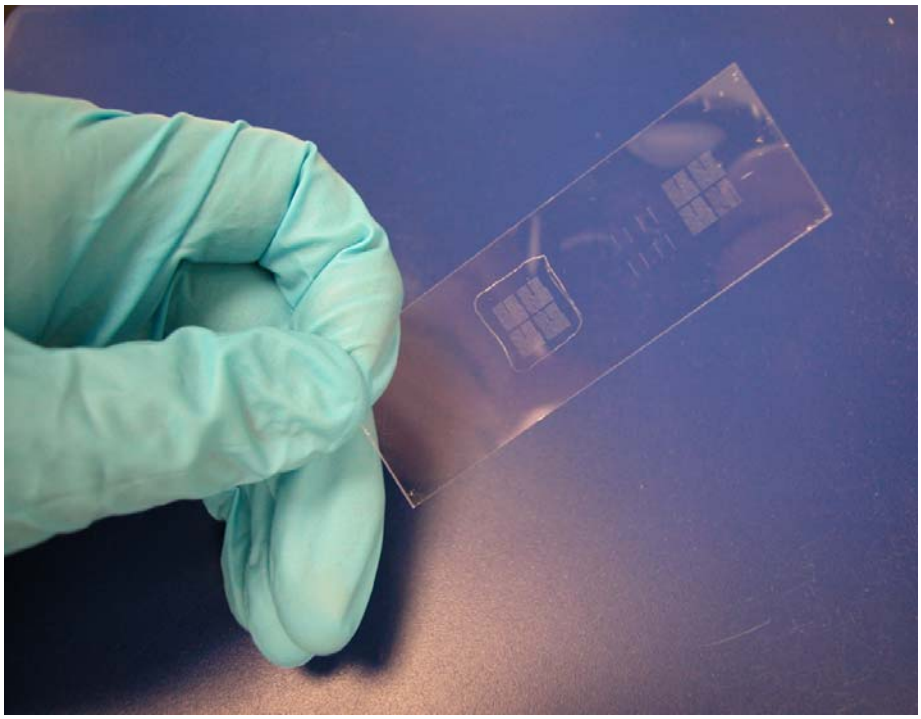


photo by Francisco Duarte

Microarray slide

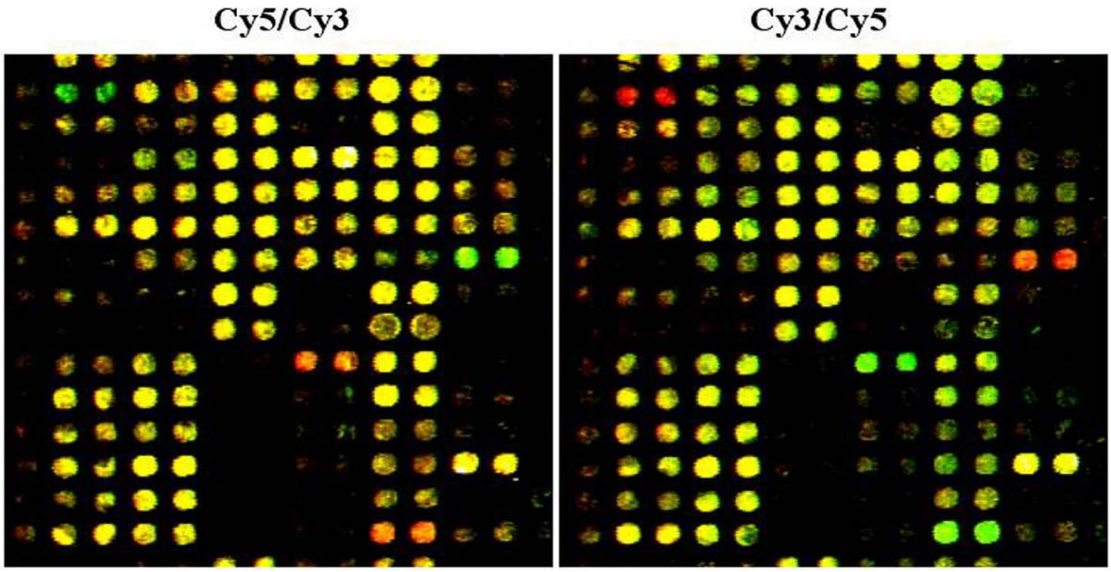


photo by Francisco Duarte

Detection of cDNA labeled with fluorescent dyes

Bioinformatics in Bioremediation : *MetaRouter*

A program developed by D. Guijas & Florencio Pazos, alma bioinformatica in collaboration with V. de Lorenzo, CNB - CSIC

Biodegradation is nature's way of recycling wastes, or breaking down organic matter into nutrients that can be used by other organisms. "Degradation" means decay, and the "bio-" prefix means that the decay is carried out by a huge assortment of bacteria, fungi, insects, worms, and other organisms that eat dead material and recycle it into new forms. By harnessing these natural forces of biodegradation, people can reduce wastes and clean up some types of environmental contaminants. Through composting, we accelerate natural biodegradation and convert organic wastes to a valuable resource. Wastewater treatment also accelerates natural forces of biodegradation, breaking down organic matter so that it will not cause pollution problems when the water is released into the environment. *Through bioremediation, microorganisms are used to clean up oil spills and other types of organic pollution.* Therefore, **Bioremediation** provides a technique for cleaning up pollution by enhancing the same biodegradation processes that occur in nature (safer, less expensive and treatment in place).

Bioremediation of a contaminated site typically works in one of two ways :

- ways are found to enhance the growth of whatever pollution-eating microbes might already be living at the contaminated site
- specialized microbes are added to degrade the contaminants (less common).

The fields of **Biodegradation** and **Bioremediation** offer many interesting and unexplored possibilities from a bioinformatics point of view. They require the integration of large amounts of data from different sources: data on chemical structures and reactivities of organic compounds; on sequencew, on structures and functions of proteins (enzymes); data about comparative genomics; environmental biology etc.

There are presently several servers providing metabolic information, among them :

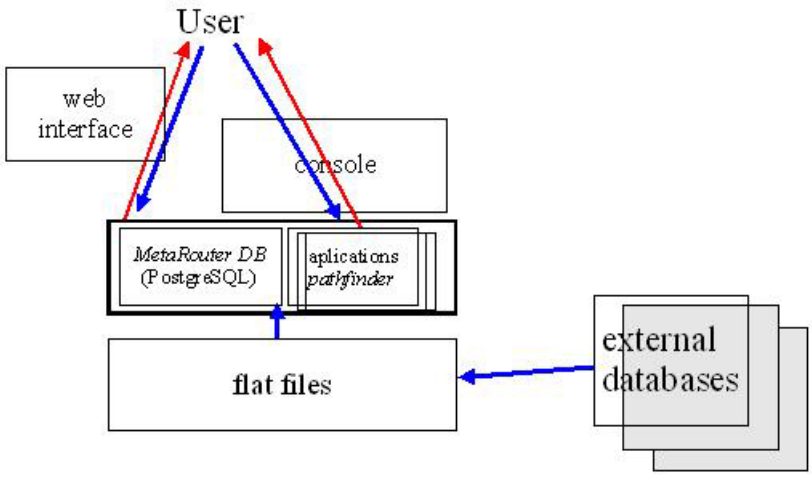
- **UM-BBD:** University of Minnesota Biocatalysis/Biodegradation Database
  http://umbbd.ahc.umn.edu/index.html
- **KEGG:** Kyoto Encyclopedia of Genes and Genomes
  http://www.genome.ad.jp/kegg/kegg.html
- **Boehringer Mannheim Biochemical Pathways** on the ExPASy server, Switzerland
  http://www.expasy.org/cgi-bin/search-biochem-index
- **Enzyme and Metabolic Pathway (EMP) Database** at Argonne National Laboratories
  http://emp.mcs.anl.gov/
- **International Society for the Study of Xenobiotics**
  http://www.issx.org/
- **Biopathways Consortium**
  http://www.biopathways.org/
- **BioCyc:** Knowledge Library of Pathway/Genome Databases
  http://biocyc.org/
- **PathDB:** Metabolic Pathways Database at NCGR

http://www.ncgr.org/pathdb/

- **Metabolic Pathway Minimaps** at Trinity College, Dublin, Ireland
  http://www.tcd.ie/Biochemistry/IUBMB-Nicholson/
- **Yeast Genome Pathways** at MIPS, Germany
  http://www.mips.biochem.mpg.de/proj/yeast/pathways/

MetaRouter is an information system for maintaining heterogeneous data related to biodegradation within a framework that allows for its administration and mining (application of methods for extracting new data). It is an application intended for laboratories which need to maintain for their own use public and private data that must be both linked internally and linked with external databases, and which also need to extract new information from it.

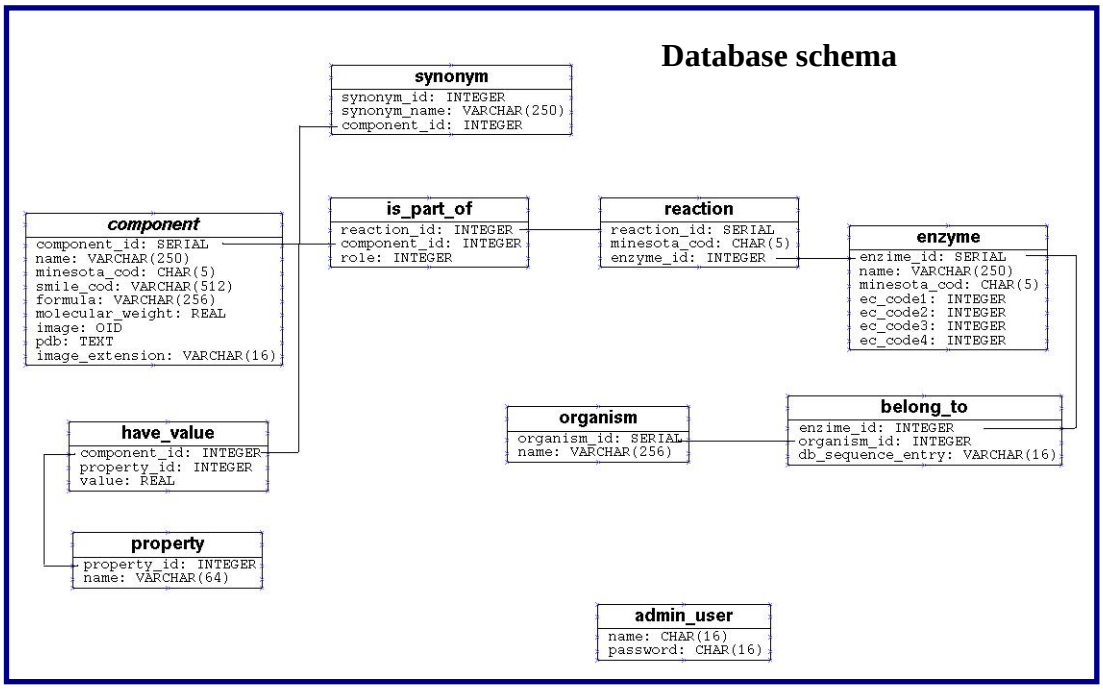
The system has an open and modular architecture adaptable to different customers. This multi-platform program, implemented in PostgreSQL (standard language for relational databases) and using SRS as an indexing system (used to connect and query molecular biology databases), works using a client/server architecture that allows the program be run on the user station or on the company server, so it can be accessed from any place in a secure way just by having a web browser.



The initial dataset is composed of 740 organic compounds (2167 synonyms), 820 reactions, 502 enzymes and 253 organisms. The subset for organic compound includes: name, synonyms, SMILES code, molecular weight, user-defined physico-chemical properties, formula, image of the chemical structure and three-dimensional structure in PDB format; for the chemical reactions: substrates, products and enzymes and; for the enzymes: name, EC code, organisms and database sequence identifiers. In all cases there are links to other databases such as :

- The University of Minnesota Biocatalysis/Biodegradation Database, **UMBBD**, (http://umbbd.ahc.umn.edu/) which is the largest source of information on biodegradation accessible by Internet.
- **ENZYME** which is a repository of information on enzymes (nomenclature, sequence, etc.) (http://www.expasy.ch/enzyme/).
- **SMILES** which is a system for coding chemical compounds as linear strings of ASCII characters. It was developed by Daylight Chemical Information Systems, Inc. (http://www.daylight.com/smiles/f_smiles.html).

The database can be summarized according to the following scheme.



MetaRouter uses two main types of inputs:

- the use of "states" (sets of compounds) to simulate an environment with a set of pollutants where a given reaction, carried out by a given bacteria, can modify one of the pollutants but not the others which "moves" the system to another "state“ (another set of compounds) where another bacteria can act, etc. For-example, this type of simulation could help with finding out which enzymes are needed to end up in the state InMet (all degraded), which are the bacteria that have them, etc.
- the addition of five properties: density, melting point (ºC), boiling point (ºC), water solubility (mg/100mL) and evaporation rate. When only qualitative solubilty information was available, the following numerical values where asigned: "insoluble": 0.0; "slightly soluble": 0.1; "soluble": 10.0 and "very soluble": 100.0. Definition of new properties and assignment of its values by the user is also allowed.

The main tool of MetaRouter is **Pathfinder**. This subprogram allows (a) the localization of pathways from an initial set of compounds to a final one and/or to the standard metabolism, (b) the selection of the pathways by length, organisms where the enzymes are present and characteristics of the implicated chemical compounds, (c) the representation of the pathways with compound name, compound image, synonyms, formula, molecular weight, SMILES code and enzyme and, hyperlinks, and, (d) color pathways according to compound properties and/or enzymatic classes.

Guided walk through *MetaRouter*

Bioinformatics in Bioremediation
MetaRouter

Guided walk through:

MetaRouter

Bioinformatics in Bioremediation
MetaRouter

Compound queries

Bioinformatics in Bioremediation
MetaRouter

Compound queries

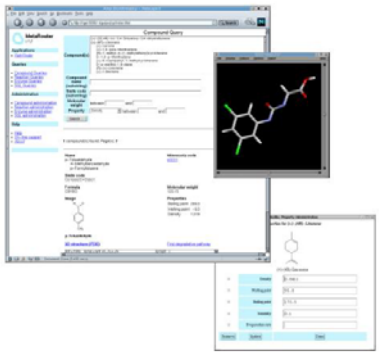
Selection of compounds by :

- Name
- Synonyms
- part of their name
- part of their smiles code
  - C=O >> comps containing carbonyl group
  - CCCCC >> comps with 5 or more linear saturated carbons
- a range of molecular weight
- a range of values of associated properties (solubility, density, etc.)

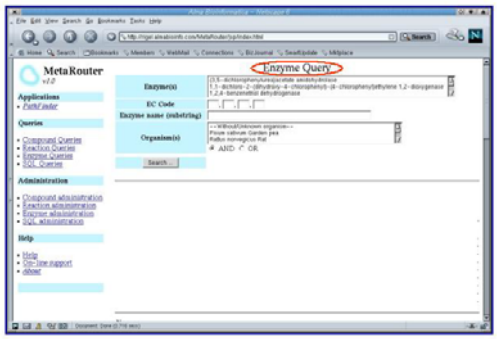
Information shown :

- name (and synonyms)
- smiles code
- formula
- image of the chemical structure
- 3D structure in PDB format
- molecular weight
- list of properties and associated values
- UMBBD code
- "Find degradative pathway"

Bioinformatics in Bioremediation
MetaRouter



Bioinformatics in Bioremediation
MetaRouter



Bioinformatics in Bioremediation
MetaRouter

Enzyme queries

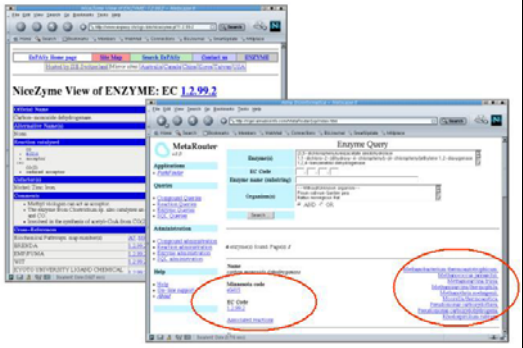
Selection of compounds by :

- enzyme name
  - part of name
- EC code
  - part of code
- organisms
- combination of some of these (i.e. EC=1 for oxidoreductases, organism=pseudomonas)

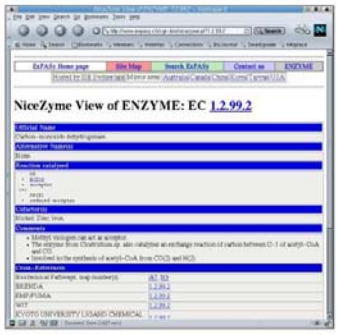
Information shown :

- enzyme name
- UMBBD code
- EC code
- organisms
- associated reactions
- links to each database

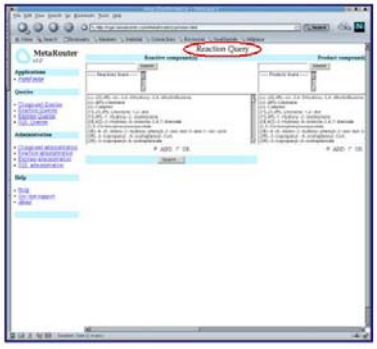
Bioinformatics in Bioremediation
MetaRouter



Bioinformatics in Bioremediation
MetaRouter



Bioinformatics in Bioremediation
MetaRouter



Bioinformatics in Bioremediation
MetaRouter

Reaction queries

Selection of compounds by :

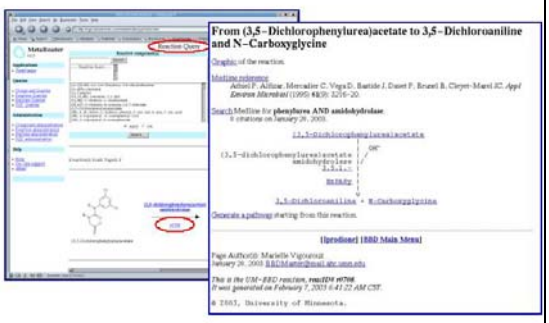
- substrate(s)
  - synonyms
  - partial names
- product(s)
  - synonyms
  - partial names
- enzyme(s)
- organism(s)

More than one substrate, product, enzyme can be selected with AND/OR (at the bottom of the list)

Information shown :

- chemical structures
  - substrate
  - products
- name of enzyme
- UMBBD code of reaction
- links to databases
  - compounds
  - enzymes
  - UMBBD page

Bioinformatics in Bioremediation
MetaRouter



Bioinformatics in Bioremediation
MetaRouter

Bioinformatics in Bioremediation
MetaRouter

SQL queries

- To directly interrogate the MetaRouter database
- Uses SQL sentences
- *Intended to expert users*
- *only selected commands are allowed (to avoid modifications of DB)*
- *SQL administration for modifying the database directly via SQL*

- Requires knowledge of database technology and SQL syntax
- but allows to carry out complex queries with just a few words.

For constructing these sentences you need to know:
- data model of the database
- name of the tables
- relations, etc. (see Database schema)
- sentences should end with a ";"

Bioinformatics in Bioremediation
MetaRouter

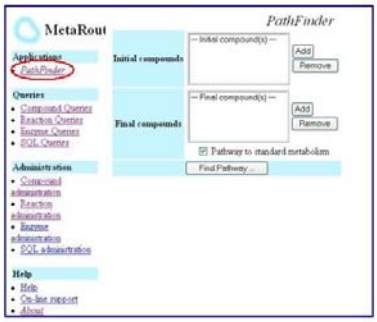
SQL queries

i.e.,

select * from enzyme where ec_code4 is not null;

this sentence will show all the enzymes in the database that are substrate specific, that is, their 4th EC code is a number (like 1.1.2.3 and not, for example 1.2.-.- that would represent a general enzymatic class):

Bioinformatics in Bioremediation
MetaRouter

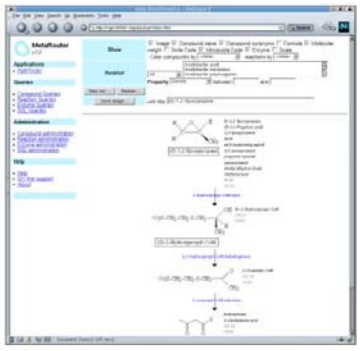


Bioinformatics in Bioremediation
MetaRouter

PathFinder

- Localization of pathways from an initial set of compounds to a final one and/or to the standard metabolism.
- Selection of the pathways by length, organisms where the enzymes are present and characteristics of the implicated chemical compounds.
- Representation of the pathways with compound name, compound image, synonyms, formula, molecular weight, SMILES code and enzyme; hyperlinked to the corresponding information for compounds, enzymes and reactions.
- Colouring pathways according with compound properties and/or enzymatic classes.

Bioinformatics in Bioremediation
MetaRouter

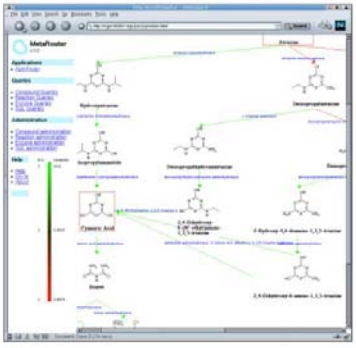


Bioinformatics in Bioremediation
MetaRouter

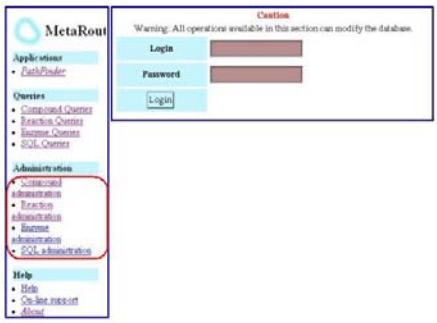
PathFinder

- Allows selection of the compound or compounds you want to degrade (initial state, *In PathFinder, a "state" is a set of compounds*).
- Selection of set of final compound(s).
- All the possible degradative pathways for this compound(s) are shown as a network of reactions.
  - all pathways
  - shortest
  - of a given organism
  - of a given range of one property
- Allows selection of elements to represent (Image, Compound name, Formula, Molecular weight, Smile Code, Minnesota Code, Enzyme and property values) the compounds (image, name, etc) are hyperlinked to the corresponding compound information pages in the database (see above), the reaction arrows are linked to the reaction information pages and the enzyme names to the enzyme information pages.

Bioinformatics in Bioremediation
MetaRouter



Bioinformatics in Bioremediation
MetaRouter



Bioinformatics in Bioremediation
MetaRouter

Administration

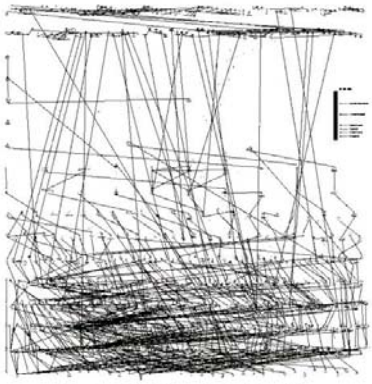
To modify the contents of (delete, add or modify) : ***– Password protected***
- Database – *Password protected*
- Compounds
- Enzymes
- Reactions

- **Modifying**: Select "item" to modify in it corresponding field (Compound, Enzyme, Reaction) by picking it from the full list or by searching by part of the name in the boxes above. On pressing "View", the information for this item is shown and you can modify it. Press "Update" at the bottom of the page to include the modifications in the database.
- **Deleting** item(s): Select one or more items and press "Delete" at the bottom of the page.
- **Inserting** item(s): Insert the information you have for the item and press "Insert" at the bottom of the page.

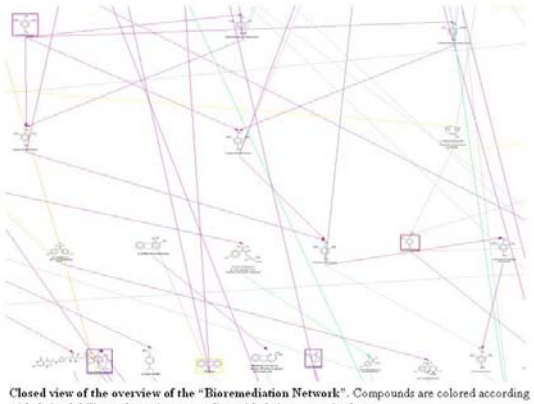
Bioinformatics in Bioremediation
MetaRouter

A practical application:

BioRemediation Network



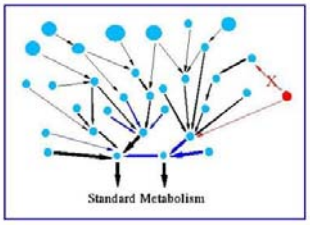
Overview of the "Bioremediation Network".



Closed view of the overview of the "Bioremediation Network". Compounds are colored according with theirsolubility and reactions according with their enzymatic class.

Bioinformatics in Bioremediation
MetaRouter

Practical Application



Bioinformatics in Bioremediation
MetaRouter

Practical Application

BioRemediation Network

- Free scale network
- Nodes closer to Standard Metabolism are more populated
- New links tend to appear bound to those most populated. It grows as free-scale networks
- Removal of those most populated nodes affect highly the stability of the network

Study carried out by :
Sito Pazos, A. Valencia & V de Lorenzo, CNB - CSIC

Bioinformatics in Bioremediation
MetaRouter

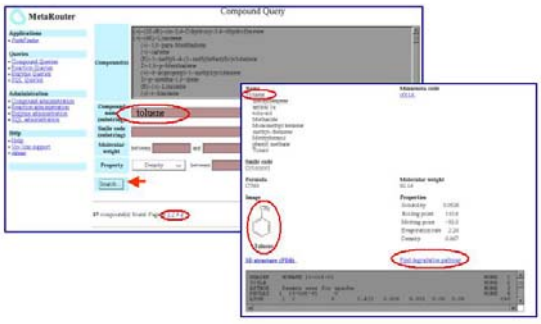
Guided Example
&
Proposed Exercises

Bioinformatics in Bioremediation
MetaRouter

Example: Analysis of possible degradative pathways for toluene

- First select **"Toluene"** in the list of initial compounds (*3rd page*). For that, you can type **"toluene"** in the search box (which will fill the search list with all the compounds containing "toluene" in their names) and then look for "Toluene" there. On pressing **"Find degradative pathway"** you will see the degradative network for toluene in a large representation. Move the scroll bars in your web browser to navigate through the representation. If you switch **off "image"** and switch **on "name" and "enzyme"** you get an easier representation with only the names of the compounds and the enzymes involved. Go back to the original representation by switching **on "images"** and switching **off "names"**. Then select **"shortest one"** and press **"Redraw"**. You see that, despite the large number of possible pathways, the shortest degradative pathway for toluene is composed of only four reactions. To see which pathways could be carried out by Pseudomonas putida select **"Show by"-"Organisms"**, **select this bacteria in the list** of organisms and then press **"Redraw"**.

Bioinformatics in Bioremediation
MetaRouter

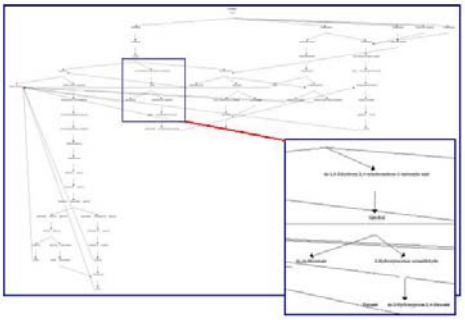


Bioinformatics in Bioremediation
MetaRouter

Example: *Analysis of possible degradative pathways for toluene*

- First select **"Toluene"** in the list of initial compounds (*3rd page*). For that, you can type **"toluene"** in the search box (which will fill the search list with all the compounds containing "toluene" in their names) and then look for "Toluene" there. On pressing **"Find degradative pathway"** you will see the degradative network for toluene in a large representation. Move the scroll bars in your web browser to navigate through the representation. If you switch **off "image"** and switch **on "name" and "enzyme"** you get an easier representation with only the names of the compounds and the enzymes involved and **Redraw**. Go back to the original representation by switching **on "images"** and switching **off "names"**. Then select **"shortest one"** and press **"Redraw"**. You see that, despite the large number of possible pathways, the shortest degradative pathway for toluene is composed of only four reactions. To see which pathways could be carried out by Pseudomonas putida select **"Show by"-"Organisms"**, **select this bacteria in the list** of organisms and then press **"Redraw"**.

Bioinformatics in Bioremediation
MetaRouter



Bioinformatics in Bioremediation
MetaRouter

Example: ***Analysis of possible degradative pathways for toluene***

- First select **"Toluene"** in the list of initial compounds (*3rd page*). For that, you can type **"toluene"** in the search box (which will fill the search list with all the compounds containing "toluene" in their names) and then look for "Toluene" there. On pressing **"Find degradative pathway"** you will see the degradative network for toluene in a large representation. Move the scroll bars in your web browser to navigate through the representation. If you switch **off "image"** and switch **on "name" and "enzyme"** you get an easier representation with only the names of the compounds and the enzymes involved. Go back to the original representation by switching **on "images"** and switching **off "names"**. Then select **"shortest one"** and press **"Redraw"**. You see that, despite the large number of possible pathways, the shortest degradative pathway for toluene is composed of only four reactions. To see which pathways could be carried out by Pseudomonas putida select **"Show by"-"Organisms"**, **select this bacteria in the list** of organisms and then press **"Redraw"**.

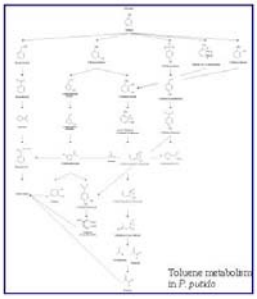
Bioinformatics in Bioremediation
MetaRouter

Bioinformatics in Bioremediation
MetaRouter

Example: ***Analysis of possible degradative pathways for toluene***

- First select **"Toluene"** in the list of initial compounds (*3rd page*). For that, you can type **"toluene"** in the search box (which will fill the search list with all the compounds containing "toluene" in their names) and then look for "Toluene" there. On pressing **"Find degradative pathway"** you will see the degradative network for toluene in a large representation. Move the scroll bars in your web browser to navigate through the representation. If you switch **off "image"** and switch **on "name" and "enzyme"** you get an easier representation with only the names of the compounds and the enzymes involved. Go back to the original representation by switching **on "images"** and switching **off "names"**. Then select **"shortest one"** and press **"Redraw"**. You see that, despite the large number of possible pathways, the shortest degradative pathway for toluene is composed of only four reactions. To see which pathways could be carried out by Pseudomonas putida select **"Show by"-"Organisms"**, **select this bacteria in the list** of organisms and then press **"Redraw"**.

Bioinformatics in Bioremediation
MetaRouter



Bioinformatics in Bioremediation
MetaRouter

Exercises:

- Let's check all the information available por **compounds** through this server on:
  - **Camphor**
  - ***o*-Xylene**
  - **TrinitroToluene** (any other compound of your interest).

-Find **reactions** on:
  - **3-methyl cathecol**
  - **between camphor and 2,5-diketocamphane**
  - **information on those enzymes involved** (any other of your interest).

- Let's inspect the possible degradative **pathways** for:
  - **trinitrotoluene**
  - **2-hydroxytoluene** (any other of your interest).
    - whole pathway (just names, no figures)
    - shortest pathway (color in function of some characteristics)
    - for some of the organisms

Protein Structure Predictions

Proteins are biomacromolecules that perform important tasks in organisms such as catalysis of biochemical reactions, transport of nutrients, recognition, and transmission of signals. Knowing the structure of a protein sequence enables us to probe the function of the protein, understand substrate and ligand interactions, devise intelligent mutagenesis and biochemical protein engineering experiments that improve specificity and stability, perform rational drug design and, design novel proteins. As a result of advances in genomic science, the gap between the number of known sequences and the number of known structures is widening rapidly. Therefore, the development of strategies able to generate probable structural models has much potential including applications which would map the functions of proteins in metabolic pathways. Simulating and generating protein folding data is one of the most fundamental unsolved problems in computational molecular biology today.

It is assumed that protein chains fold up into a unique 3D structure, that is at the free-energy minimum (although it is known that chaperones play a role in some folding pathways); that many different sequences can adopt the same basic fold; that the main driving force is the need to pack hydrophobic residues into the interior of the molecule; that analysis of the underlying chemistry shows that this is only possible if the protein forms regular patterns of a macroscopic substructure called secondary structure. Thus, all information about the native structure of a protein is coded in the amino acid sequence within a specific solution environment. In practice, however, this code cannot be deciphered since not all parameters are known, the energy differences between native and unfolded proteins are extremelly small and the high computing time required for such complex analysis. That is why actually, predictions are based on the combination of empirical and statistical methods. CASP experiments carried out recently have demonstrated that structure predictions from the polypeptide sequence only are not possible yet.

There are three major theoretical methods for predicting the structure of proteins: comparative modelling, fold recognition, and *ab initio* prediction.

- **Ab initio prediction**. These methods will mimic the protein folding process by computing the molecular dynamics based on our knowledge of the physical laws and micro-environment in cells. However, we do not yet have a complete understanding of the driving forces behind protein folding. These methods exhibit two major subproblems: (1) sampling the conformational space of the protein well so that a significant number of native-like conformations are generated, and (2) designing a discriminatory/scoring/energy function that will distinguish between native and non-native conformations in this sample.
- **Comparative or "homology" modelling**. Structure is more conserved than sequence. Protein pairs with more than 30% pairwise sequence identity (for alignment length > 80%) have homologous 3D structures (the essential fold of the two proteins is identical, additional loop regions may vary). So, the main idea consists of modelling the protein of unknown structure on the template of a

sequence homologue of known structure. However, this method only allows to predict 10-30% of all protein sequences.

- **Fold recognition or "threading"**. There is evidence that most pairs of proteins with similar structure are remote homologues with less than 25% pairwise sequence identity. The basic idea is to thread the sequence of the query into the known structure target and to evaluate the fitness of sequence for structure by some kind of environment-based or knowledge-based potential.
- **Accurate prediction for 1D aspects of 3D structure**. If not remote homologue can be detected for our query sequence, the prediction problem has to be simplified. Using the diversity of information in current databases, it is possible to make very accurate 1D predictions for each of the residues within its sequence context and for regular patterns of macroscopic substructures. Several prediction services are readily available for analysing secondary structure, solvent accessibility, protein compactness, location and topology of transmembrane helices, and the location of helices for the special class of coiled-coil proteins, phosphorilation sites, signal peptides etc… Most of these servers implement neuronal networks and evolutive information to improve the predictions since evolution takes place on a 3D structural level and not at sequence level.

Ab initio predictions

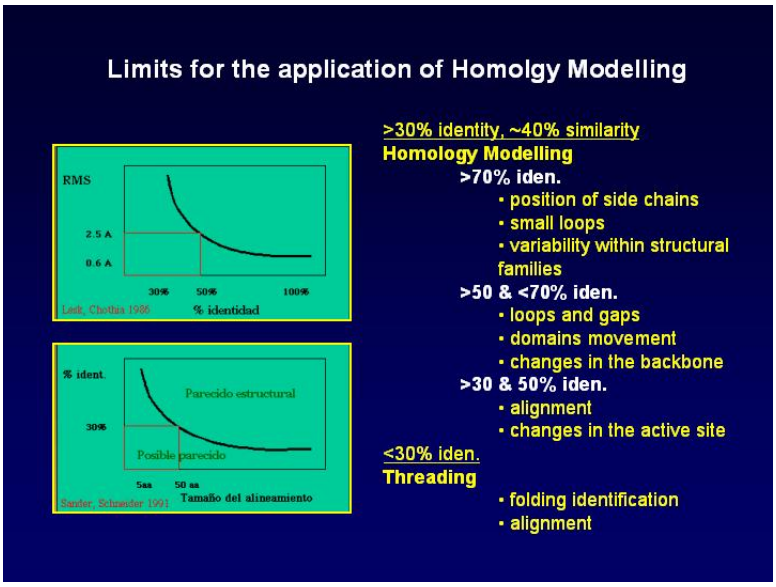
The *ab initio* approach tries to understand how the three-dimensional structure of proteins is attained and to deduce the three-dimensional structure given the sequence. The biggest challenge for these methods can be broken down into two components: devising a scoring function that can distinguish between correct (native or native-like) structures from incorrect (non-native) ones, and searching a method to explore the conformational space. Currently there does not exist a reliable and general scoring function that can always drive a search to a native fold, and there is no reliable and general search method that can sample the conformation space adequately to guarantee a significant fraction of near-natives.

Some methods for *ab initio* prediction include Molecular Dynamics (MD) simulations of proteins and protein-substrate complexes provide a detailed and dynamic picture of the nature of inter-atomic interactions with regards to protein structure and function; Monte Carlo (MC) simulations that do not use forces but rather compare energies via the use of Boltzmann probabilities; Genetic Algorithms which tries to improve on the sampling and the convergence of MC approaches, and exhaustive and semi-exhaustive lattice-based studies which are based on using a crude/approximate fold representation (such as two residues per lattice point) followed by exploring all or large amounts of conformational space given the crude representation.

Comparative modelling

Comparative modelling exploits the fact that evolutionarily related proteins with similar sequences, as measured by the percentage of identical residues at each position based on an optimal structural superposition, have similar structures. The similarity of

structures is very high in the so-called ``core regions", which typically are comprised of a framework of secondary structure elements such as alpha-helices and beta-sheets. Loop regions connect these secondary structures and generally vary even in pairs of homologous structures with a high degree of sequence similarity. These theoretical approach provides the molecular biologists with "low-resolution" models which hold enough essential information about the spatial arrangement of important residues to guide the design of experiments.



The process of building a comparative model is conceptually straightforward. First, comparative protein modelling requires at least one sequence of known 3D-structure with significant similarity to the target sequence. To determine if this "homology modelling method" can be applied to a particular sequence, one compares the target sequence with a sequences of Brookhaven Protein Data Bank (PDB), using programs such as FastA, BLAST or PsiBlast.

Protein Structure Databases:

- *CATH : http://www.biochem.ucl.ac.uk/bsm/cath/*
- *GenBank : http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html*
- *GeneCensus : http://bioinfo.mbb.yale.edu/genome*
- *ModBase : http://guitar.rockefeller.edu/modbase/*
- *PDB : http://www.rcsb.org/pdb/*
- *Presage : http://presage.berkeley.edu/*
- *PSI : http://www.structuralgenomics.org/*
- *SCOP : http://scop.mrc-lmb.cam.ac.uk/scop/*
- *TrEMBL : http://srs.ebi.ac.uk/*

Sequence Search Servers:

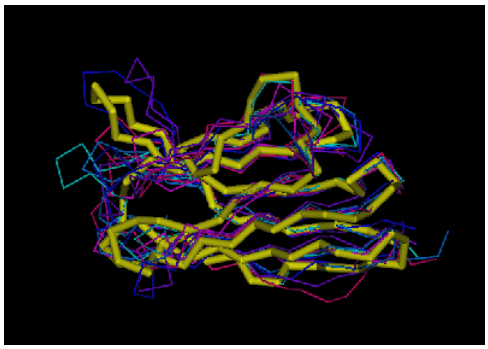
- *BCM : http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html*
- *BLAST2 : http://www.ncbi.nlm.nih.gov/gorf/bl2.html*
- *BLOCK MAKER : http://blocks.fhcrc.org/blocks/blockmkr/make_blocks.html*
- *ClustalW : http://www2.ebi.ac.uk/clustalw/*
- *FASTA3 : http://www2.ebi.ac.uk/fasta3/*

The above procedure might allow the selection of several suitable templates for a given target sequence to be used in the modelling process. The best template structure - the one with the highest sequence similarity to the target - will serve as the *reference*. The selected templates will be superimposed onto it in 3D. The 3D structurally corrected multiple sequence alignment is achieved by using the best-scoring diagonals obtained by SIM. This superposition can then be optimised by maximising the number of Ca pairs in the common core while minimising their relative mean square deviation. Each residue of the reference structure is then aligned with a residue from every other available template structure of their Ca atoms. This generates a structurally corrected multiple sequence alignment.

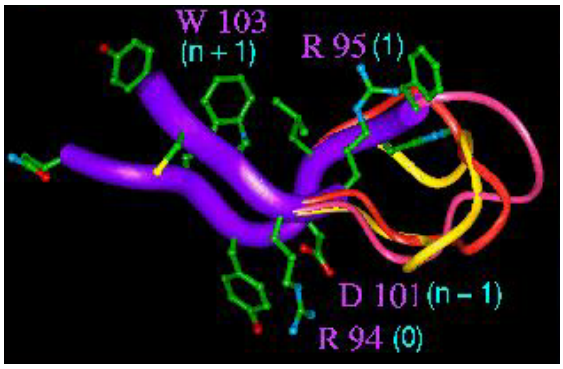
Homology Modelling Servers:

- ***SWISS-MODEL***. *http://www.expasy.ch/swissmod/SWISS-MODEL.html*
- ***CPHmodels***. *http://www.cbs.dtu.dk/services/CPHmodels/*
- ***SDC1***. *http://cl.sdsc.edu/hm.html*
- ***Modeller***. *http://guitar.rockefeller.edu/modeller/modeller.html*
- ***3D-JIGSAW***. *http://www.bmm.icnet.uk/servers/3djigsaw/*

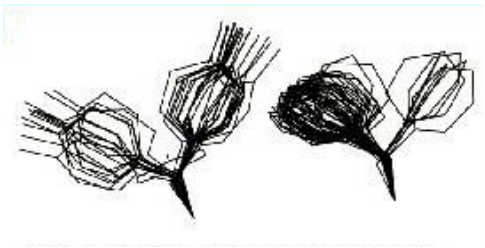
The next step is the construction of a framework which is computed by averaging the position of each atom in the target sequence, based on the location of the corresponding atoms in the template. When more than one template is available, the relative contribution, or weight, of each structure is determined by its local degree of sequence identity with the target sequence.



Then, loops for which no structural information was available in the template structures and therefore are not defined, must added in a further step. Using a "spare part" algorithm, one searches for fragments which could be accommodated onto the framework from the PDB entries. Each loop is defined by its length and its "stems", namely the alpha carbon (Ca) atom co-ordinates of the four residues preceding and following the loop. The fragments fitting our loop definition (our gap) will be extracted from a library of pentapeptide backbone fragments derived from the PDB entries. These fragments are then fitted to overlapping runs of five Ca atoms of the target model. The co-ordinates of each central tripeptide are then averaged for each target backbone atom (N, C, O) and added to the model.



For many of the protein side chains there is no structural information available in the templates. Therefore these cannot be built during the framework generation and must be added later. The number of side chains that need to be built is dictated by the degree of sequence identity between target and template sequences. To this end one uses a table of the most probable rotamers for each amino acid side chain depending on their backbone conformation. All the allowed rotamers of the residues missing from the structure are analysed to see if they are acceptable by a van der Waals exclusion test. The most favoured rotamer is added to the model. The atoms defining the angles of incomplete side chains can be used to restrict the choice of rotamers to those fitting these angles.



Idealisation of bond geometry and removal of unfavourable non-bonded contacts and atomic clashes (bumps) can be performed by energy minimisation with force fields such as CHARMM, AMBER or GROMOS. The refinement of a primary model should be performed softly since experience has shown that energy minimisation (or molecular dynamics) usually modifies the model structure away from the control structure. It is thus necessary to keep the number of minimisation steps to a minimum and in soft conditions. Constraining the positions of selected atoms (such as Ca, or using a B-factor based function in each residue generally helps avoiding excessive structural drift during force field computations.

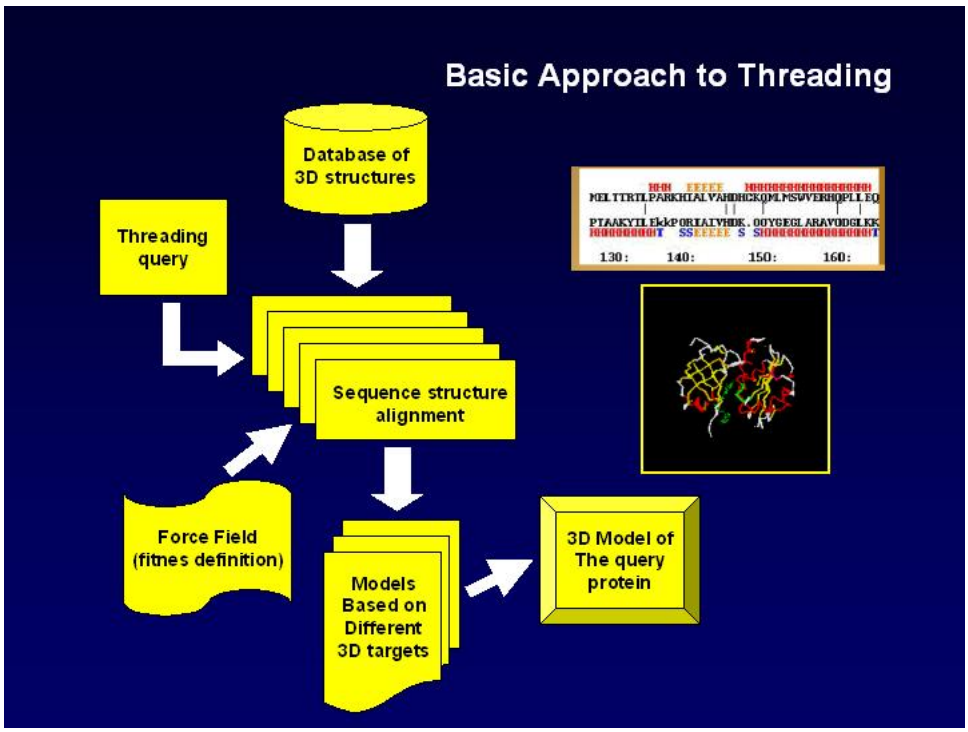
Fold recognition or “threading”

Methods of protein fold recognition attempt to detect similarities between protein 3D structure that are not accompanied by any significant sequence similarity (less than 25% homology). All these methods are based on the finding of protein folds that are compatible with a particular sequence. Unlike “homology methods”, threading methods take advantage of the extra information made available by 3D structure information. In fact, rather than predicting how a sequence will fold, they predict how

well a fold will fit to the query sequence. These methods use libraries of known three-dimensional structures including pairwise atom contacts and solvation terms and, by using a scoring function assess the fit of the query sequence to a given fold. These methods can be extremely elaborate such as those involving double dynamic programming, dynamic programming with frozen approximation, Gibbs Sampling using a database of ``threading'' cores, and branch and bound heuristics, or as simple as using sophisticated sequence alignment methods such as Hidden Markov Models. Despite initially promising results, methods of fold recognition are not always accurate. CASP competition has shown that even when the methods were successful, alignments of sequence on to protein 3D structure were usually incorrect implying that comparative modelling performed using such models would be inaccurate, that careful human insight, a knowledge of known structures, secondary structure predictions and thoughts about the function of the target sequences are a powerful aid during fold recognition.

The results suggest that one should use caution when using these methods. In spite of this, the methods remain very useful.

These methods fit the following scheme:



Fold Recognition (threading) Servers:

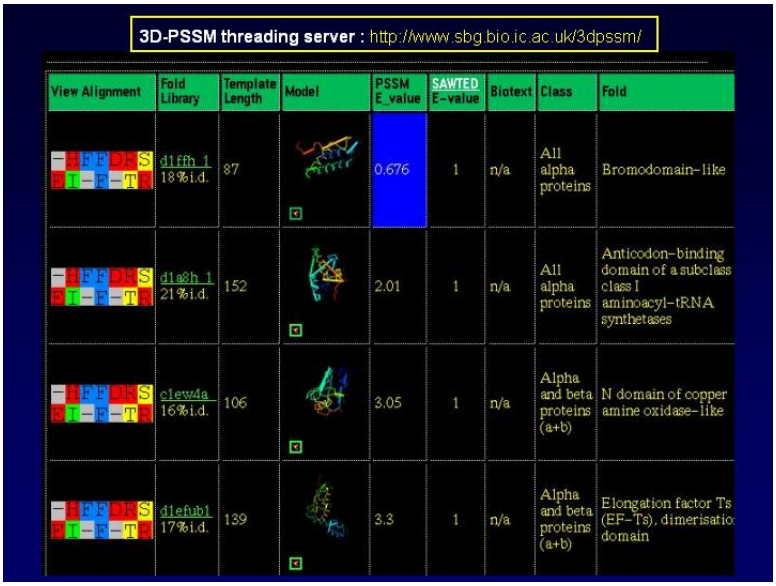
- ***3D-PSSM***. *http://www.bmm.icnet.uk/~3dpssm/*
- ***Threader2***. *http://insulin.brunel.ac.uk/threader/threader.html*
- ***TOPITS (PHDthreader)***. *http://cubic.bioc.columbia.edu/predictprotein/*
- ***SAM-T99***. *http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99-model-library-search.html*
- ***FUGUE***. *http://www-cryst.bioc.cam.ac.uk/~fugue/*
- ***bioinbgu***. *http://www.cs.bgu.ac.il/~bioinbgu/form.html*

- ***123+***. *http://www-lmmb.ncifcrf.gov/~nicka/123D+.html*
- ***Threadlize.*** *Interactive visualization and combination of threading models. http://www.cnb.uam.es/~pazos/threadlize/*

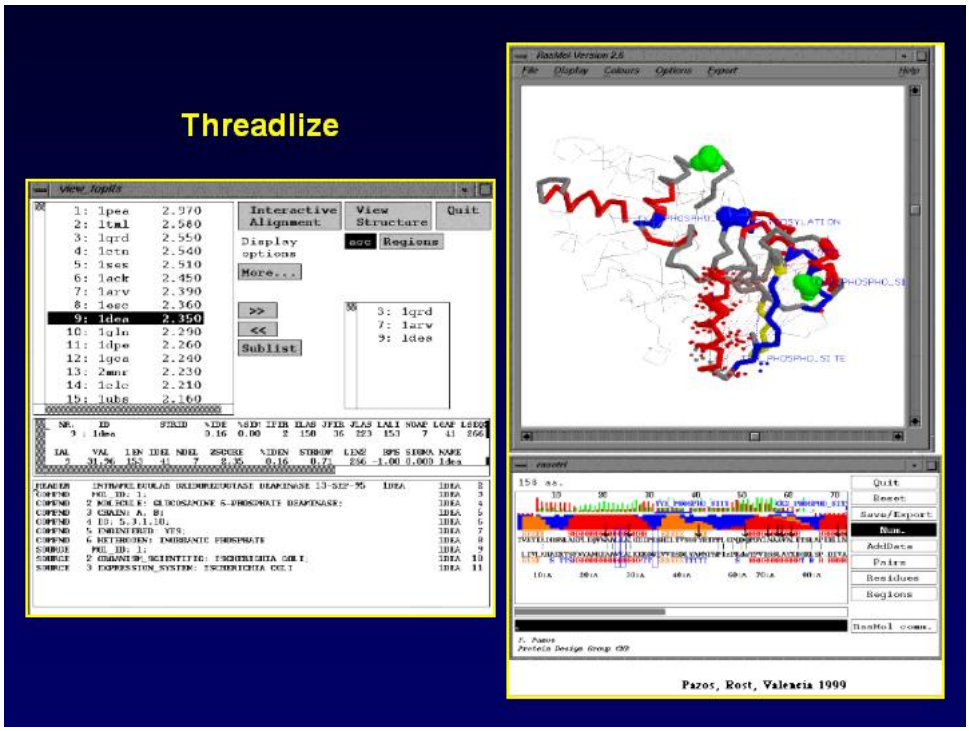
The practical approach is to run several of these methods, and run each of them for several of the sequence homologues found for our query sequence. (That is, a blast search should be performed and several of the top homologues should be sent to several of these threading servers.) Then, a consensus picture of the likely fold can be build up. In addition, it should be considered that:

- the correct fold may not be at the top of the list, but that it is likely to be in the top 10 scoring folds
- knowledge of the function of the query protein can provide insight to search for remote homologues and can help with protein fold recognition using humans rather than computers
- alignments from these servers can be used as a starting point but the best alignment of sequence on to tertiary structure is still likely to come from careful human intervention

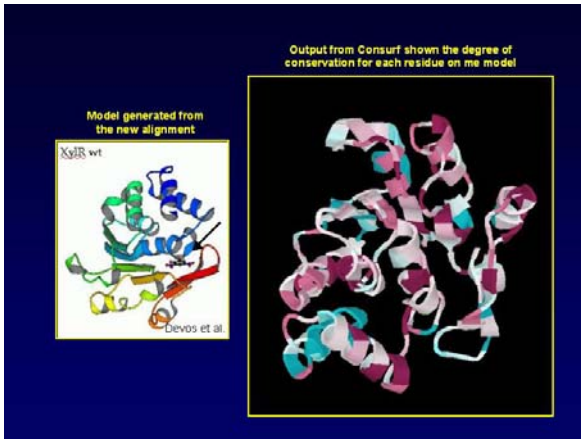
Shown below is the output of 3D-PSSM threader server that includes the sequence alignment, the backbone proposed for the model fitting each of the hits with a goodness score and, some information on the folding class and function of each of the hits.



Shown below is the analysis of output from one of the threader servers using "threadlize". This tool allows one to evaluate and edit the alignments returned by the server, so that it can be improved by addition of information from secondary structure, solvent accessibility, etc. The server also has a viewer so you can also analyse specific residues that can be important for the protein.

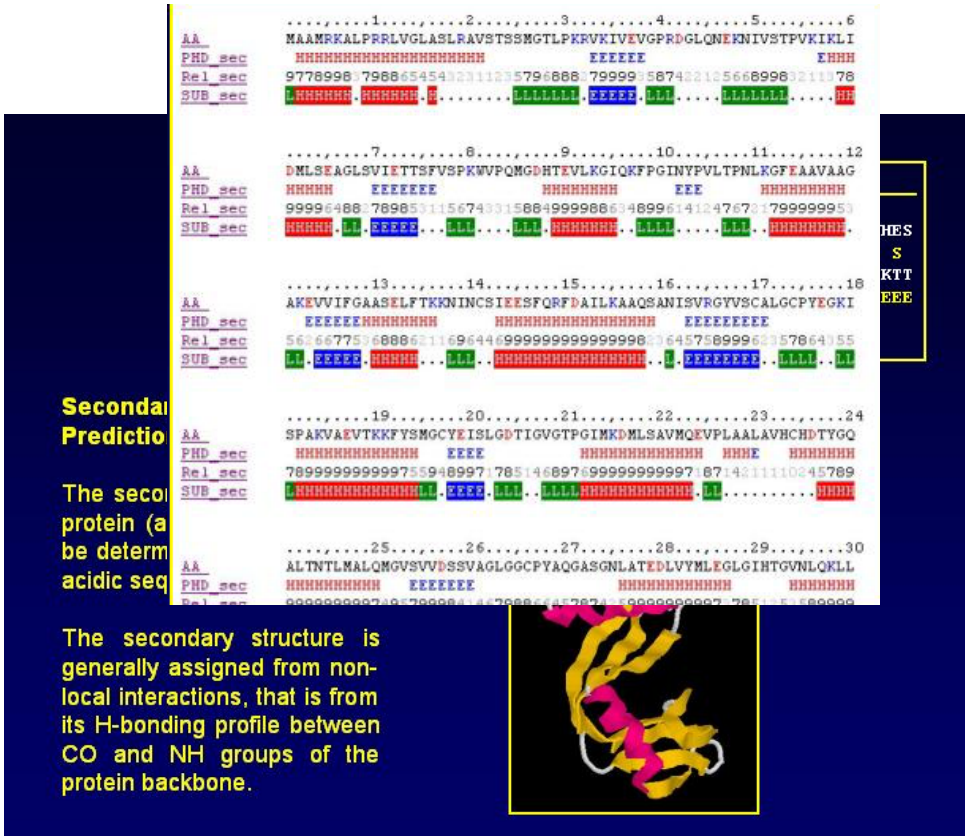


Shown below is the final proposed structure for the query sequence.



Accurate prediction for 1D aspects of 3D structure

1D-characteristics are those that can be represented by a single simple code that can be assigned to each of the amino acids of the séquense. (B. Rost). Secondary structure for instance can be represented by 1) associated value: H, B, L for helices, betas or loops respectively); 2) accessibility (e: exposed; b: buried); 3) hydrophobicity, etc.. These properties are useful for the prediction of 3D-structures.



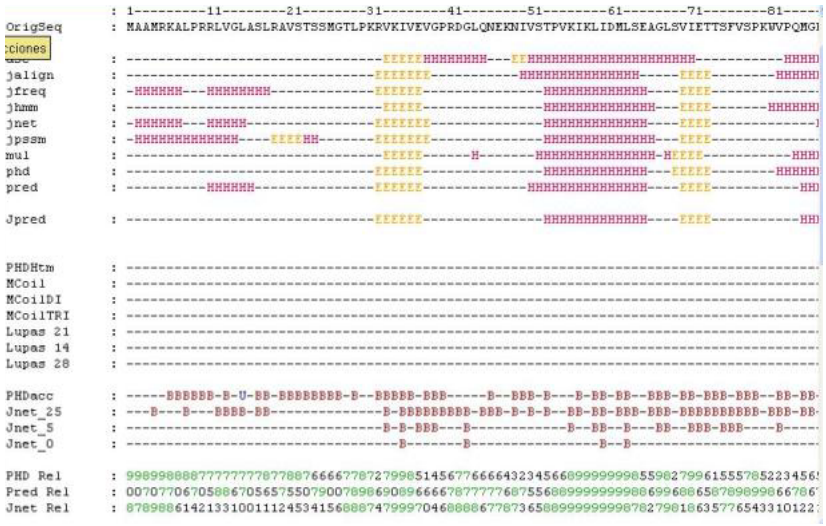
Several **physico-chemical properties** can be extracted, such as: hydrophobicity, polarity, etc. These properties can be of utility for the prediction of structural characteristics. ***ProtScale*** within the *ExPASy Tools* takes a sequence as imput and returns up to 54 different scalar parmeters in a graphical way (*http://www.expasy.ch/cgi-bin/protscale.pl*).

Secondary structure prediction (alpha-beta-loop) methods use neuronal networks and/or other algorithms trained by a set of proteins of known structure. And they also include evolutionary information from múltiple alignments.

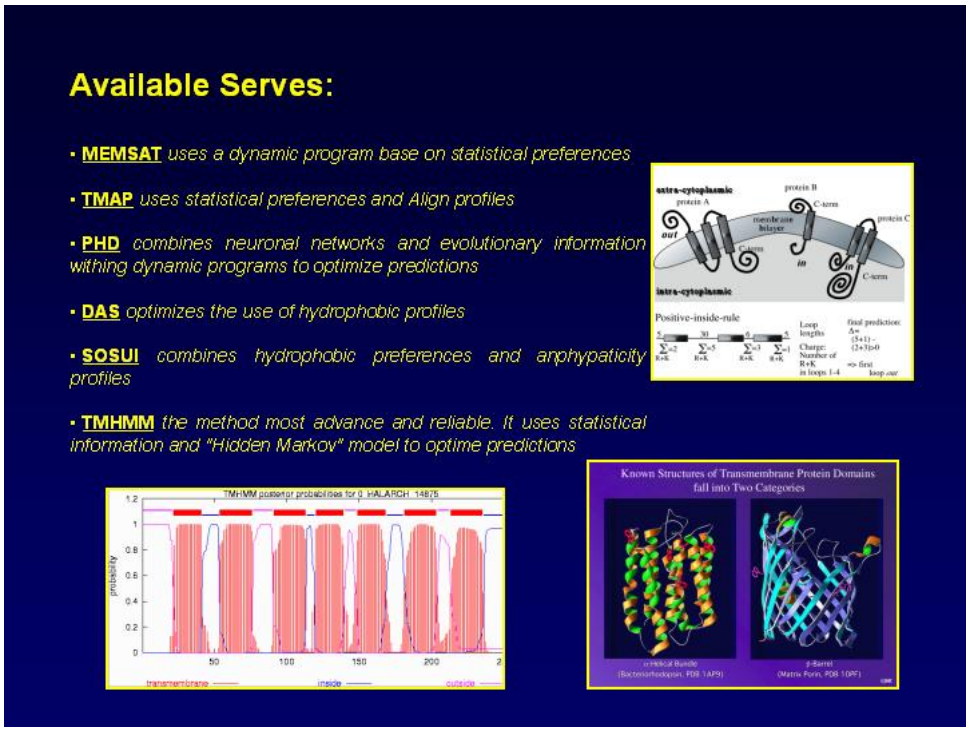
- *Secondary Structure Prediction.* *Prediction of the secondary structure state for the residues along the sequence.*
  - ***PHDsec***. *http://cubic.bioc.columbia.edu/predictprotein/*
  - ***PROF*** *(former DSC).* *http://www.aber.ac.uk/~phiwww/prof/*
  - ***Predator***. *http://www.embl-heidelberg.de/predator/doc_www.html*
  - ***PSI-pred***. *http://www.psipred.net/*
  - ***Agadir***. *Prediction of alpha-helical content of small peptides.* *http://www.embl-heidelberg.de/Services/serrano/agadir/agadir-start.html*
  - ***Consensus methods***. *These servers run different secondary structure prediction methods and combine the results.*
    - ***JPred***.*http://jura.ebi.ac.uk:8888/*
    - ***NPS@***.*http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html*

Accessibility to the solvent is of interest for modelling a sequence. The most detailed method evaluating the volume exposure to the solvent by each residue has been developed by Connolly and is implemented in DSSP. (Output accessibility < 16% *"buried"* or ≥ 16% *"exposed"*).

- *Solvent Accessibility Prediction.* *Prediction of the solvent accessibility status of the residues along the sequence. Most of the methods for secondary structure prediction include also solvent accessibility prediction.* ***PHDsec***. *http://cubic.bioc.columbia.edu/predictprotein/*
- ***PROFphd,*** PHD y PROFphd (*from PredictProtein*) apply neuronal networks and multiple Align information. These servers provide numeric values for the accessibility (matrix with values of 0, 1, 4, 9, 16, 25, 36, 49, 64, 81).
- ***JPred2***, uses PsiBlast profiles as input for the neuronal networks and returns two state-values (buried or exposed).



Transmembrane proteins. One of the biggest challenges of proteomics is the determination of the structure of transmembrane proteins (difficult to crystalize and determine by NMR). There are two main groups of transmembrane proteins: those getting alpha helices in the membrane and, those forming pores made off beta barrels (type porines). *So far there are not publicly available servers to determine the second type of transmembrane proteins (because of the lack of experimental information). However, the situation is quite different for the first type.* The 3D structure of these proteins can be determined knowning the precise position of its helices just by checking all the possible conformations.



- *Transmembrane Helix Prediction. Prediction of residues forming part of membrane spanning segments. Some of the servers predict also the orientation of the segment (N-terminal intracellular or cytoplasmic).*
  - ***PHDhtm/PHDtopology**. http://cubic.bioc.columbia.edu/predictprotein/*
  - ***HMMTOP**. http://www.enzim.hu/hmmtop/*
  - ***TMpred**. http://www.ch.embnet.org/software/TMPRED_form.html*
  - ***TopPred 2**. http://www.sbc.su.se/~erikw/toppred2/*

Coiled-coil is a "double alpha-helix": two helices with a characteristic distribution of residues to establish a dimeric interaction forming a "superhelix". This program is based on neuronal networks trained with known coiled-coil structures.

- *Coiled-coil Prediction. Prediction of coiled-coil regions. A "coiled-coil" is a super-helix formed by two of three coiled helices.*
  - ***COILS**. Lupas's method. http://www.ch.embnet.org/software/COILS_form.html*
  - ***Multicoil**. Berger's method. It also predicts the multimerization state of the coil (dimeric or trimeric). http://nightingale.lcs.mit.edu/cgi-bin/multicoil*
- *Other Sequence Features. Localization of sequence tracks known to play biologically important roles.*
  - ***NetOglyc**. Prediction of O-glycosilation sites. http://www.cbs.dtu.dk/services/NetOGlyc-2.0/abstract.html*
  - ***PSORT**. Prediction of protein sorting signals and localization sites. http://psort.nibb.ac.jp/*
  - ***SignalP**. Prediction of signal peptide cleavage sites. http://www.cbs.dtu.dk/services/SignalP/*

- *ChloroP. Prediction of chloroplast transit peptides.* *http://www.cbs.dtu.dk/services/ChloroP/*
- ***MITOPROT** . Prediction of mitochondrial targeting sequences.* *http://www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter*
- ***big-PI**. GPI Modification Site Prediction.* *http://mendel.imp.univie.ac.at/gpi/*
- ***NetPhos**. Prediction of Ser, Thr and Tyr phosphorylation sites in eukaryotic proteins.* *http://www.cbs.dtu.dk/services/NetPhos/*

Sequence Motifs. Motifs are small polypeptide frements (5-10) associated with a specific function. Structural Domains are segments of polypeptide chains with specific structural and/or functional characteristics that differentiate them from others.

- *Domains and Motifs. Localization of domains and motifs common to known protein families.*
  - ***Prosite**. Biologically significant sites, patterns and profiles.* *http://www.expasy.ch/prosite/*
  - ***Pfam**. Collection of Hidden Markov Models of sequence alignments.* *http://www.sanger.ac.uk/Pfam/*
  - ***ProDom**. Collection of profiles derived from psi-blast alignments.* *http://protein.toulouse.inra.fr/prodom.html*

Exercise:

- Locate the PDB file with the 3D structure.
  - Search a sequence: BLAST against PDB.
- Using this tool calculate the physico-chemical properties. ExPASy ProtScale.
- Calculate secondary structure, solvent accessibility, transmembrane proteins, etc… PredictProtein.
- C*oiled-coil regions*. Coils.
- Sequence Motifs

>Protein

```
MSKPQPIAAANWKCNGSQQSLSELIDLFNSTSINHDVQCVVASTFVHLAMTKERLSHPKF
VIAAQNAIAKSGAFTGEVSLPILKDFGVNWIVLGHSERRAYYGETNEIVADKVAAAVAS
GFMVIACIGETLQERESGRTAVVVLTQIAAIAKKLKKADWAKVVIAYEPVWAIGTGKVA
TPQQAQEAHALIRSWVSSKIGADVAGELRILYGGSVNGKNARTLYQQRDVNGFLVGGASLKPEFVDIIKA
TQ
```