# Antecedentes

## Existe promiscuidad a nivel enzima, a nivel familia e inclusive a nivel ruta de metabolismo especializado.

El estudio del metabolismo puede abordarse a varios niveles: genes, enzimas, familias enzimáticas, rutas metabólicas, organismos e inclusive comunidades. De la misma forma si se entiende la promiscuidad como funciones metabólicas alternativas puede observarse promiscuidad a distintos niveles, desde un mismo gen que presenta splicing alternativo . la variación es la materia prima de la evolucion. la promiscuidad enzimática podemos estudiarla en esos mismos. tres de ellos son: i) en una enzima de un organismo en particular, ii) en una familia de enzimas homólogas iii) en rutas de metabolismo especializado. En este trabajo se discierne entre las preguntas ¿es esta familia promiscua?, ¿Dentro de esta familia promiscua, es promiscua la enzima perteneciente a este organismo?, ¿Es este cluster biosintético promiscuo?. El primer problema consiste en ubicar cuáles familias enzimáticas tienen al menos una enzima promiscua, esta es una propiedad de las familias. El segundo problema consiste en identifcar individualmente enzimas promiscuas en una familia promiscuas, esta es una propiedad de las enzimas no de las familias. Y finalmente si entendemos un cluster biosintético como un conjunto de enzimas que reciben un sustrato, podemos generalizar la noción de promiscuidad a rutas biosintéticas de metabolismo especializado.

## Antecedentes conceptuales

Para mejorar nuestro entendimiento del fenomeno, ademas de la comparacion de secuencias es necesario integrar otros elementos de analisis. Se debe notar que es practicamente imposible decir que una enzima no es promiscua ya que para ello se deberian haber descartado todos los posibles sustratos. Sin embargo, para el estudio de cambios en promiscuidad se han detectado como elementos relevantes a los cambios en vecindad genomica, los cambios en flexibilidad durante la dinamica molecular, la perdida de genes centrales, y finalmente a las expansiones genicas dentro de un grupo taxonomico. Estos elementos tienen en comun que reflejan un cambio en alguna propiedad genomica o biofisica, de lo que se deriva que el buscar cambios en la promiscuidad de una enzima resulta mas factible que la busqueda intrinseca de promiscuidad, a lo que aspiran los metodos basados en comparaciones de secuencias.

Se ha intentado identificar enzimas promiscuas, a nivel de secuencia, sin pasar por experimentacion mediante aprendizaje maquina. Estos enfoques son incapaces de identificar una familia promiscua si no se conoce previamente al menos un miembro promiscuo de ella. Por otra parte, en el problema de los miembros se presentan dificultades cuando la identidad de secuencia es alta, e.g. en la familia PriA-HisA se sabe que la enzima HisA de E. coli no es promiscua, pero PriA de Streptomyces coelicolor si lo es.

## El establecimiento de un marco de conservación permite distinguir cambios

Bacterias de la misma especie tienen genomas distintos. Para poder establecer diferencias en el número de copias debemos trabajar en la construcción de un marco filogenético consistente que permita ordenar inclusive organismos de la misma especie. Para ello pueden utilizarse el core genome el total de genes conservado de cada genoma en un grupo taxonómico. [@van\_der\_veen\_metaphor\_2014] piloplan Para distinguir cambios primero se tiene que establecer un marco de conservación.

## La genómica comparativa como herramienta en la distinción de miembros promiscuos  
Evomining es una plataforma bioinformatica pensada para la busqueda de expansiones de familias genicas de metabolismo central. Desarrollarla en combinacion con algoritmos de busqueda de cambios en la vecindad genomica la haran una plataforma ideal para abordar el problema de las familias, proporcionando una solucion a la dificultad de no tener conocimiento previo de un miembro promiscuo en la familia investigada. Respecto al problema de los miembros, se propone explorar variaciones en vecindad genomica, flujo genico y dinamica molecular, como candidatos a reflejar la variacion en promiscuidad. Finalmente, he detectado que a pesar de que pruebas in vivo son mas sensibles a niveles bajos de promiscuidad que mediciones in vitro, esta ultima suele ser la unica estudiada. In vivo, la metabolomica aplicada en genes biosinteticos de productos naturales ha ayudado en la identificacion de sustratos, por lo que esta tecnica podria ayudar a revelar el nuevo sustrato de una enzima en la que se sospecha ganancia de promiscuidad. En resumen, el objetivo de mi trabajo sera abordar los dos problemas de promiscuidad considerando la diferenciacion in vitro e in vivo tomando como modelo biologico el phylum Actinobacteria, un grupo de bacterias reconocido por su diversidad metabolica donde se ha probado la existencia de promiscuidad enzimatica.

Evomining es una plataforma bioinformatica pensada para la identificacion de productos naturales que tiene entre sus exitos la identificacion de la biosintesis de arsenolipidos [@cruz-morales\_phylogenomic\_2016]. La busqueda de productos naturales cuenta entre sus premisas que estos se producen en vecindades genomicas llamadas clusters y que ademas clusters cercanos (ya sea en contenido genico o en la secuencia de sus componentes), exploran variaciones metabolicas, es decir sus enzimas catalizan reacciones sobre sustratos parecidos aunque no identicos [@cruz-morales\_phylogenomic\_2016,@medema\_computational\_2015]. La base de evomining es que las enzimas de metabolismo secundario son expansiones distantes de enzimas de rutas centrales, lo que da idea de la quimica que realizan dichas expansiones dejando por identificar el sustrato sobre el que trabajan. La primera version de evomining cuenta con 200 genomas de Actinobacteria, una base de datos de secuencias de enzimas de productos naturales y otra base de datos de secuencias de enzimas de rutas centrales curada a mano. Evomining esta ligada con el problema de la promiscuidad porque en estas familias expandidas ya sea por duplicacion o por transferencia horizontal, las expansiones pueden retener la funcion quimica de las rutas centrales y viceversa, la funcion quimica expandida suele estar presente antes de la duplicacion.

Si se combinara evomining con la premisa de que vecindades distintas son marcadoras de funciones quimicas distintas, al encontrar una familia expandida con vecindades genomicas diferentes se podria solventar la deficiencia de otros metodos bioinformaticos consistente en que para identificar familias promiscuas se debe conocer previamente un miembro promiscuo de la misma. (Fig 4) Asi pues al combinar evomining con herramientas de vecindad genomica tanto de comparacion como de visualizacion estaremos mejorando su funcionalidad en la identificacion de familias promiscuas.

## La genómica comparativa como herramienta en la priorización de clusters promiscuos

La promiscuidad nos interesa por su produccion de variantes. Si bien en rutas centrales rescata la función en metaboismo secundario crea nuevas variantes moleculares que permiten adaptación, de hecho pangenomas grandes correlacioan con aparición de nuevas funciones enzimáticas. Considero que el concepto de promiscuidad puede ser extendido a un nuevo niivel. Promiscuidad enzimatica, promiscuida de familia, promiscuidad de cluster. En rutas centrales robustez, pero en metabolismo secundario platicidad. Pangenoma abierto cerrado, Aqui proponemos organizar los clusters y finalmente la medida de su apertura. variantes de clusters producen nuevos compuestos, ya sea por promiscuidad enzimática en un core conservado o por variación en la presencia / ausencia de genes. Un cluster recibe los mismos sustratos y los transforma en diferentes productos.

### Expansion y contextos genomicos como herramienta de anotacion funcional

La diversidad enzimatica existente es el resultado de un proceso de expansion, mutacion y seleccion que se ha desarrollado durante el transcurso de la historia evolutiva [@khersonsky\_enzyme\_2010, @pearson\_prehistoric\_2012]. Existe evidencia de que cierto grado de promiscuidad o divergencia funcional precede a la duplicacion genica [@hughes\_evolution\_1994]. Por este motivo detectar expansiones ya sea duplicaciones o transferencias horizontales [@treangen\_horizontal\_2011], puede ser un buen punto de partida para determinar divergencia funcional y promiscuidad. No todas las expansiones denotan cambio de funcion enzimatica, algunas pueden ser meros accidentes, sin embargo dado que la funcion de una enzima suele estar relacionada con sus vecinos [@overbeek\_use\_1999,@zhao\_prediction\_2014,@zhao\_discovery\_2013, @verdel-aranda\_molecular\_2015], una expansion en una vecindad genomica diferente de la tradicional sera un referente de adquisicion de una nueva funcion y entonces un indicador de existencia previa de promiscuidad.

La funcion de una enzima es un concepto jerarquico, dependiente de la filogenia de un organismo [@szklarczyk\_string\_2015]. Para sistematizar el estudio de contextos y vecindades genomicas se desarrollo Search Tool for the Retrieval of Interacting Genes/Proteins STRING [@snel\_string\_2000], que cuenta con una anotacion de ortologia jerarquica y consistente, realizada en 2000 organismos en cuyo marco interacciones de proteinas con implicaciones funcionales son predichas tanto de novo por informacion genomica de co-ocurrencia como por mineria de datos en articulos publicados. STRING es una base de datos, y como tal no permite agregar nuevos genomas para su analisis. Sus 2000 organismos incluyen especies tanto bacterianas como eucariotas. Al existir tanta diversidad, los genomas disponibles para un genero o clase especificos son escasos, p. g. de los mas de 300 genomas disponibles de Streptomyces solo 24 estan incluidos.

Para resolver la baja cobertura de STRING hacia ciertos grupos taxonomicos se pueden desarrollar scripts de vecindad genomica utilizando RAST (Rapid Annotation using Subsystem Technology); un servicio interactivo de anotacion automatica de genomas de bacterias y arqueas [@aziz\_rast\_2008,@overbeek\_seed\_2014] donde la funcion de cada gen se asigna de acuerdo a conocimiento previo de subsecuencias de organismos cercanos filogeneticamente, cuando es posible se incluye en un subsistema metabolico. Estamos en una era de explosion de datos genomicos, proximamente se espera contar con millones de genomas bacterianos incluso provenientes de bacterias no cultivables, por ello los algoritmos deben ser constantemente optimizados a los nuevos volumenes de datos [@medema\_computational\_2015]. Ante esta expectativa seria muy util desarrollar algoritmos de analisis genomico que sean de codigo libre o al menos interactivos para que cada laboratorio pueda personalizarlos para sus propios genomas.

Finalmente, no solo la vecindad genomica inmediata puede ser utilizada como distintivo en la busqueda de promiscuidad, diferencias en el contexto genomico en genes relacionados con una enzima promiscua, sin importar su ubicacion dentro del genoma tambien pueden ser relevantes para la perdida o ganancia de funcion quimica [@noda-garcia\_evolution\_2013], (Juarez Vazquez et al 2015).

### Metodos bioinformaticos

Al evaluar PROMISE [@carbonell\_molecular\_2010] en un set de datos de la familia HisA/PriA [@noda-garcia\_insights\_2015] obtuve que su mejor desempeño es con huella molecular de tamaño 6, donde clasifica correctamente casi todas las no promiscuas, (HisA) pero no sucede lo mismo con la familia PriA donde tiene exito en 16 de 45 casos. Al aplicar el mismo tamaño de huella a 9 miembros promiscuos de la familia IlvC no consigue predecir correctamente ninguno de ellos. Por lo menos para estas familias el conjunto de entrenamiento o los descriptores no son suficientes para la anotacion de promiscuidad.

### Contexto y vecindades genomicas

En 2012 fueron analizados 102 genomas de 29 familias de Actinobacteria [@noda\_estudio\_2012]. sugiriendo que al menos en *Corynebacteria* el contexto y la vecindad genomica incidian en la sub-funcionalizacion de PriA en subHisA [@noda-garcia\_evolution\_2013]. Respecto a IlvC, otra familia involucrada en la sintesis de aminoacidos fue estudiada y caracterizada bioquimicamente en 1 Corynebacterium y 8 Streptomyces [@verdel-aranda\_molecular\_2015]. Para ampliar estos resultados, utilizando la anotacion de RAST y una generalizacion de la definicion de vecindad de STRING, se diseño un algoritmo para identificar vecindades similares asi como uno de visualizacion de contexto, ambos disponibles como software libre en github [nselem/perlas](https://github.com/nselem/perlas) .

El algoritmo de clasificacion de vecindades permite agruparlas en clusters y calificar estos clusters segun su conservacion dado un grupo de bacterias. La definicion de vecindad y similitud de vecindad esta descrita posteriormente en los metodos. El algoritmo fue aplicado a la familia IlvC en 290 Streptomyces resultando 9 clusters [Datos](http://148.247.230.43/nselem/CONTEXTS/REL_St275/ilvC/Contextos.php) entre los mas poblados el primero cuenta con 279 elementos, otro con 9 elemento y dos mas con 7 miembros (Fig 3), resultados experimentals son congruentes con que existe divergencia funcional entre miembros de clusters distintos [@verdel-aranda\_molecular\_2015]

Natural products genomic era[@harvey\_re-emergence\_2015]

## Estudio de una familia PriA

### Caracterizacion in vivo

Algunas enzimas PriA no han mostrado promiscuidad in vitro pero si in vivo ya que sobreviven en un medio sin triptofano, es decir in vivo complementan la funcion trpF. Para la construccion de cepas de Streptomyces con variantes no nativas de priA minimizando la modificacion genomica y el efecto de sobreexpresion, se planea utilizar E. coli como intermediario para realizar seleccion por auxotrofia. Se cuenta con un conjunto de plasmidos para transformar a E. coli asi como con las mutantes sencillas de E. coli para trpF y hisA que permiten realizar seleccion por auxotrofias. Ademas tenemos una coleccion de cepas nativas de Streptomyces asi como un mutante de PriA de S. coelicolor. Se optimizo una reaccion de PCR para la amplificacion de un segmento de DNA de S. coelicolor que contiene a priA.

### Caracterizacion bioquimica in vitro.

De la familia PriA y sus subfamilias se han caracterizado bioquimicamente miembros selectos de Actinomycetaceae, Bifidobacteriaceae, Micrococcaceae, Acidimicrobiaceae, Corynebacterium, Mycobacteriaceae, Streptomycetaceae, Camera (provenientes de metagenoma), reconstrucciones ancestrales, 80 mutantes de Corynebacterium, y 2 mutantes de Camera mediante cineticas enzimaticas para calcular las constantes Kcat,Km. El genero Streptomyces, el que cuenta con mayor cantidad de genomas disponibles representa una oportunidad muy poco explotada de explorar la influencia del contexto y la vecindad genomicas en secuencias de PriA (Tabla 3, Figura 5).

### Modelado de dinamica molecular

La dinamica es un metodo que permite hacer simulaciones de particulas que sirve para obtener informacion de propiedades macroscopicas de un conjunto de atomos [@petrenko\_molecular\_2001,@kukol\_molecular\_2008]. Es util en el marco de mi proyecto porque permite la exploracion del espacio conformacional, y se ha visto que este esta relacionado con la actividad de la enzima [@sikosek\_biophysics\_2014], ademas dado un conformero permite verificar su estabilidad. Resuelve la ecuacion de movimiento de Newton con base a una configuracion inicial, las fuerzas interatomicas como los enlaces covalentes, las fuerzas de Van der Waals y la carga de las particulas[@campbell\_biophysical\_2012]. Entonces para generar una simulacion de dinamica molecular, debe contarse con una estructura como punto de partida, ya sea esta cristalografica o modelada de novo o por homologia. El laboratorio de bioinformatica y biofisica computacional ha desarrollado un protocolo de generacion de modelos homologos estructurales y dinamicas moleculares (Carrillo-Tripp et al 2015 in prep); con este pipeline se han generado dos estructuras de Camera [@noda-garcia\_insights\_2015], 30 estructuras y dinamicas de miembros de Actinobacteriaceae y Bifidobacteriaceae (Vazquez-Juarez et al in prep.) y finalmente una estructura de subHisA de Corynebacterium diphteriae. En la familia Streptomyces, interesante debido a su variacion en contexto genomico y en mediciones in vitro aun no se modelan dinamicas moleculares aunque 40 estructuras por homologia estan en proceso.

En un estudio de subHisA [@noda\_estudio\_2012] se utilizo el metodo de dinamica molecular y se comparo el numero de conformeros entre miembros de subHisA y PriA, resultando mayor el de PriA como corresponde a una enzima promiscua. El estudio sobre la relacion dinamica-flexibilidad de $\noindent\beta$-lactamasas utiliza replica exchange, una variacion de dinamica molecular que corre replicas en paralelo a distintas temperaturas [@bai\_replica\_2006]. Una desventaja de este metodo es que por el costo computacional de las replicas agregar explicitamente otras moleculas a la simulacion como el solvente no es posible en tiempo razonable. Una vez generadas las dinamicas moleculares se procedera a calcular tanto el numero de conformeros como el indice de flexibilidad dsi [@zou\_evolution\_2015]. Se esta desarrollando PEDB, promiscuous enzyme database, una base datos genomicos, evolutivos, bioquimicos y estructurales y de metabolismo de PriA en Actinobacteria donde se procedera al analisis de los mismos (<http://148.247.230.43/nselem/PHP/queries.html>).

En conclusion la promiscuidad enzimatica es un fenomeno complejo debido a multiples causas. Existe una gran variedad de estudios con enfoques puntuales sobre aspectos estructurales, dinamicos y evolutivos de familias de enzimas promiscuas, sin embargo hasta ahora no se han reportado trabajos multidisciplinarios que involucren a todas las partes involucradas (Fig 6)