INAIL

2023

BIG DATA ANALYSIS E INTELLIGENZA ARTIFICIALE:
STRUMENTI INAIL A SUPPORTO DEI RICERCATORI NELLA GESTIONE DELLA MOLTITUDINE DEI DATI - OMICI

PREMESSA

Tematica di approfondimento riguardante la "Big data analysis in ambito Biotecnologico" che si inserisce negli obiettivi della raccolta prevista dal Fact sheet dal titolo: "Applicazioni biotecnologiche gli aspetti normativi e i progetti inail", pubblicato nel 2022.
In questa scheda informativa sono inizialmente presentate le nuove tecniche NGS (Next Generation Sequencing) che consentono ai ricercatori di studiare ed elaborare una grande quantità di dati depositati in repository pubblici che sono potenzialmente ricchi di informazioni su eventi cellulari. Tali dati possono essere utilizzati per dare risposte a quesiti biologici non ancora studiati, consentendo ai ricercatori di fare nuove scoperte, estraendo e rianalizzando, con nuove domande biologiche, set di dati pubblici.
Successivamente, vengono presentati gli strumenti Inail a supporto dei ricercatori nella gestione della moltitudine dei dati - omici, la banca dati molecolare BiTdata e l'applicazione di Intelligenza Artificiale (IA), di cui vengono schematizzate le loro potenzialità, i flussi nell'analisi dei dati e alcune elaborazioni significative.

INTRODUZIONE

Lo sviluppo delle nuove tecnologie correlate allo studio delle scienze "omiche" ha apportato una rivoluzione nel modo di fare ricerca; si è passati infatti da un approccio basato su ipotesi ad un approccio basato sui dati, che a volte possono rispondere molto più velocemente a quesiti biologici ancora aperti.
Con il termine bioinformatica, coniato nel 1970 da Paulien Hogeweg e Ben Hesper, si intende descrivere «lo studio dei processi informatici nei sistemi biotici»; la bioinformatica è un campo interdisciplinare che sviluppa metodi e strumenti software per estrarre conoscenza dal dato biologico.
Per sequenziamento genetico di nuova generazione (NGS), noto anche come sequenziamento ad alto rendimento, si intende l'insieme delle tecnologie di sequenziamento degli acidi nucleici che hanno in comune la capacità di sequenziare, in parallelo, milioni di frammenti di DNA. È quindi il termine generico utilizzato per descrivere una serie di diverse tecnologie, che consentono di sequenziare il DNA e l'RNA ovvero il sequenziamento del genoma, il risequenziamento del genoma, il profiling del trascrittoma (RNA-Seq), le interazioni DNA-proteina (sequenziamento del ChIP) e la caratterizzazione dell'epigenoma. La necessità di elaborare, archiviare ed analizzare l'enorme quantità di informazioni ottenute dalle tecnologie NGS (Big Data) ha portato allo sviluppo di molte soluzioni per l'elaborazione di dati, basate sul Deep Learning (classe di algoritmi Machine Learning che utilizzano diversi livelli di astrazione per dare un significato ai dati) e di molti database open source resi disponibili per i ricercatori di tutto il mondo (Figura 1).

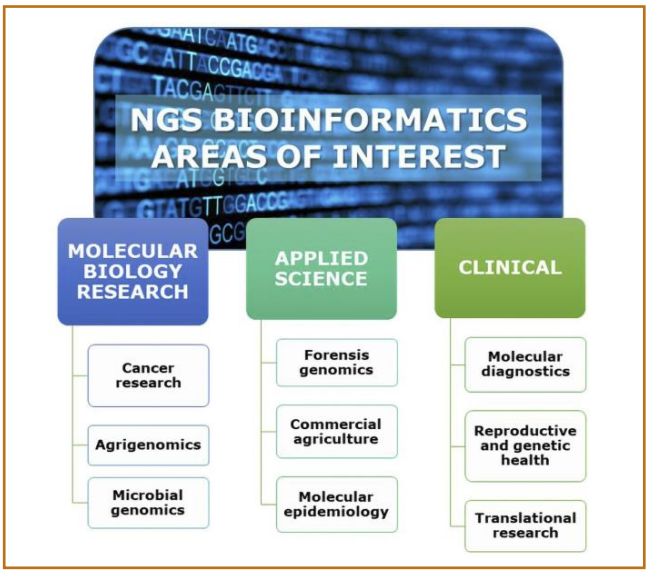

Figura 1. Aree di interesse per il sequenziamento di nuova generazione (NGS).

Le più importanti applicazioni NGS includono: 1) lo studio della regolazione dell'espressione genica, comprese modifiche epigenetiche, interazioni tra proteine e sequenze regolatorie, predizione di varianti di splicing dell'RNA messaggero (mRNA); 2) l'analisi del genoma e la ricerca sugli SNP, nelle regioni codificanti e non codificanti del genoma, la previsione della struttura delle proteine; 3) la diagnosi precoce di patologie attraverso la ricerca e monitoraggio dei biomarcatori.

Ogni disegno sperimentale può indirizzare molteplici domande di tipo biologico.

Le tecniche HTS (High Throughput Sequencing) applicate allo studio dei cambiamenti di espressione dei geni, rappresentano uno straordinario strumento per velocizzare l'identificazione di cambiamenti patologici a livello molecolare, ancor prima che si manifestino segni clinici dell'eventuale insorgenza di patologie, e un booster alla possibilità di identificazione di nuovi potenziali biomarcatori. Il dato di trascrittomica, infatti, riflette le variazioni che il trascrittoma subisce tra cellula e cellula o tessuto e tessuto, in seguito a mutamenti delle condizioni in cui la cellula si trova, e quindi rappresenta un utile strumento di valutazione del comportamento molecolare di cellule e tessuti, in condizioni fisiologiche e/o patologiche.

Il riutilizzo dei dati può aiutare a formare un ecosistema di conoscenza biologica

Grandi volumi di dati di trascrittomica ottenuti, dispo-

nibili in database pubblici, possono essere rianalizzati successivamente per dare risposte a quesiti biologici non ancora studiati, perché non conosciuti o non ottenibili a causa della disponibilità limitata di dati al momento in cui è stata condotta una prima analisi. Questa nuova disponibilità di dati biomedici consente quindi di fare nuove scoperte semplicemente estraendo e rianalizzando set di dati depositati, da soli o insieme ad altri set dati, poiché ricchi di informazioni su eventi cellulari, e quindi ottenere nuove informazioni riguardo la correlazione tra esposizione a sostanze di diversa origine e geni differenzialmente espressi.
Tali approcci sono, ad esempio, particolarmente vantaggiosi nello studio dell'esposizione a vari fattori ambientali, correlati e non a luoghi di lavoro. L'analisi potrebbe infatti evidenziare quali mutazioni o alterazioni dell'epigenoma siano più frequenti in individui esposti a specifiche condizioni ambientali, fornendo anche potenziali biomarker che potrebbero rivelarsi utili per l'attività di screening e diagnosi precoce in categorie a rischio (Figura 2).

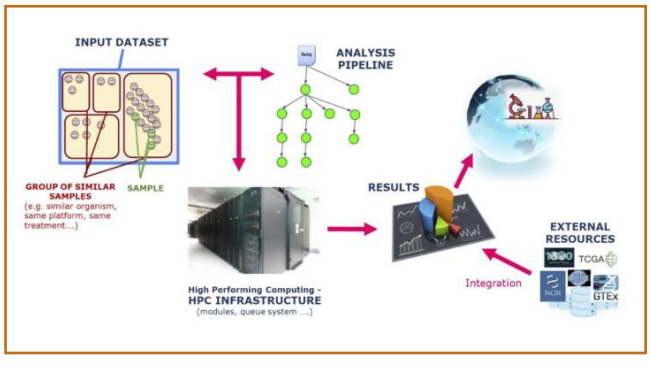

Figura 2. Data Management workflow. Attraverso nuove pipeline di lavoro, è possibile identificare processi biologici convergenti, vie di signaling a seguito di esposizione occupazionale e/o ambientale e possibili nuovi geni candidati biomarcatori.

Da questa esperienza nasce il nuovo Progetto di ricerca scientifica Inail in corso dal titolo: "Big data analysis e intelligenza artificiale nell'elaborazione di dati molecolari derivanti da esposizione ambientale a xenobiotici" (Figura 3).

Figura 3. Il nuovo progetto di ricerca scientifica Inail 2023-2025. Responsabile Scientifico dott.ssa Miriam Zanellato. L'esposizione ambientale e professionale cronica a contaminanti e agenti cancerogeni costituisce un fattore di rischio per lo sviluppo di numerose patologie. Lo studio dell'espressione differenziale di geni permette di correlarne i cambiamenti allo sviluppo di diverse patologie. Il progetto è ancora in corso e si sta procedendo con l'elaborazione dei dati di trascrittomica pubblicati e disponibili al fine di evidenziare come l'RNA non codificante possa contribuire allo sviluppo e al differenziamento, in modo particolare, delle patologie polmonari.

Le conoscenze che derivano dagli studi di epigenomica e trascrittomica, attraverso le nuove tecniche di NGS rivestono un grande potenziale come fattori informativi delle interazioni che intercorrono tra il genoma e l'ambiente e lo sviluppo di eventuali effetti negativi sulla salute associati all'esposizione a xenobiotici ambientali ed occupazionali. Si sta assistendo infatti ad un crescente interesse riguardo la loro possibile inclusione nel processo di valutazione del rischio (Figura 4).

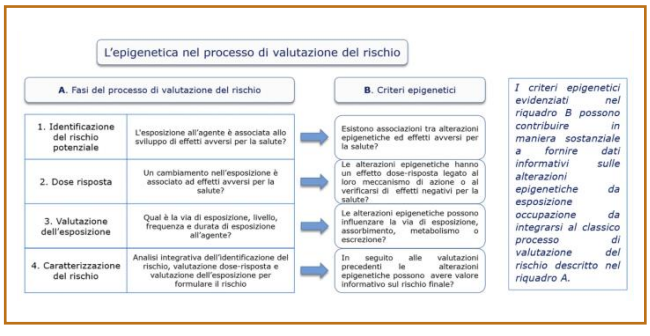

Figura 4. L'epigenetica nel processo di valutazione del rischio.

BITDATA, POTENZIALITÀ E FLUSSI PER L'ANALISI DEI DATI

In riferimento a quanto già evidenziato nel precedente fact sheet allo scopo, quindi, di individuare set di dati rilevanti a fini prevenzionistici, il dit in collaborazione con Inail-Dcod e l'Università degli Studi "Sapienza" di Roma ha realizzato una Banca dati molecolare INAIL, denominata BiTdata, in assonanza ai Big data. Si tratta infatti di "dati molecolari BioTecnologici" che fungono quasi da "Biglietto Tecnologico" dell'esposizione occupazionale. La banca dati BiTdata è consultabile nella sezione dedicata alle attività di Ricerca e Innovazione tecnologica del sito Inail. BiTdata prende in esame Piattaforme Informatiche Internazionali, che mettono a disposizione dataset completi dei principali cambiamenti genomici in seguito ad esposizione occupazionale ad agenti fisici chimici e biologici. È progettata, quindi, allo scopo di individuare set di dati rilevanti a fini della prevenzione, rendendo fruibili e accessibili i dati depositati nei "repository" pubblici; consentirebbe di effettuare "meta-analisi", ovvero analisi di campioni con caratteristiche biologiche comparabili, che rappresenta una delle più importanti sfide della bioinformatica. Gli studi del trascrittoma, inoltre, combinati con tecniche di data mining, possono fornire nuove informazioni sulla patogenesi di numerose patologie e possono contribuire all'identificazione di nuovi biomarcatori candidati con potenziale valore clinico.
Quindi, nell'ambito delle attività del progetto sopra descritto, un ulteriore stream di ricerca ha previsto l'impiego di tecnologie cognitive, in collaborazione con il partner tecnologico Inail DCOD e IBM Italia, per facilitare e velocizzare la raccolta delle informazioni per la nuova Banca dati molecolare.
L'attività di raccolta dei dati della BiTdata è stata automatizzata, attraverso tecnologie di intelligenza artificiale, e viene rinnovata con cadenza regolare al fine di mantenere costantemente aggiornato il database (Figure 5-6).

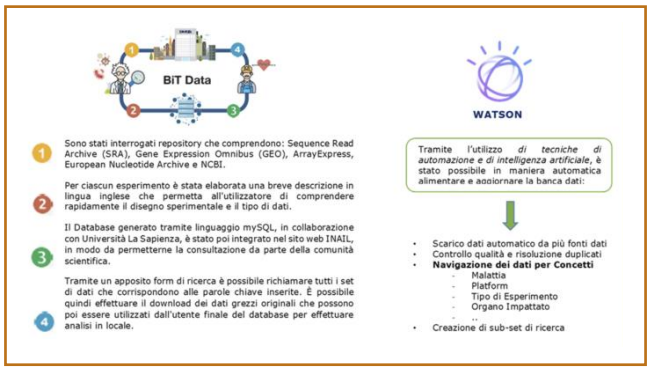

Figura 5. Sono state schematizzate in figura le funzionalità di Bitdata della piattaforma di IA, strumenti Inail a supporto dei ricercatori nella gestione della moltitudine dei dati – omici.

IL FLUSSO DELL'APPLICATIVO DI INTELLIGENZA ARTIFICIALE (IA)

In particolare, la piattaforma sviluppata è in grado di interrogare automaticamente i repository genomici scelti, che mettono a disposizione in modalità "open-source" i dati e a seconda di una lista di sostanze d'interesse da monitorare, preleva le informazioni rilevanti di ogni esperimento. La lista può essere aggiornata e rivista da un'interfaccia, in questo modo è stato possibile recentemente aggiungere anche SARS-CoV-2 fra le richieste di interrogazione. La richiesta dei dati avviene una volta al mese, con una priorità differente fra i vari repository, dando precedenza a GEO Dataset, il più fornito fra quelli individuati e cercando successivamente di andare per differenza, in modo da non rischiare di salvare dei duplicati.
Per ogni esperimento l'applicazione salva non solo l'abstract, ma una serie di dati ed informazioni che il repository mette a disposizione, come il titolo dell'esperimento, l'anno, la tecnica di sequenziamento e vari altri. Successivamente degli algoritmi di AI appositamente sviluppati, agiscono nella descrizione dell'esperimento per andare ad estrarre altre informazioni rilevanti, come la presenza di termini derivanti dal dizionario MeSH (Medical Subject Headings), biomarcatori specifici, il tipo di esperimento se in vitro o in vivo. Una lista completa delle informazioni recuperate direttamente dai repository o rilevate dall'utilizzo di tecniche di intelligenza artificiale è la seguente:

- Agente;
- Fonte dati;
- Tipo di piattaforma;
- Data di pubblicazione;
- Tipo di esperimento;
- Biomarcatore;
- Terminologia MeSH;
- Time Course;
- Numero di campioni;
- Tipo di soggetto;
- Tipo di studio.

Ogni informazione o dato aggiuntivo rispetto a quelli forniti e presenti nei repository, abilita delle nuove analisi e permette di creare in maniera più precisa dei subset di studi specifici. Si potrà ad esempio chiedere all'applicazione di visualizzare tutti gli esperimenti dell'anno 2020, con tecnica di sequenziamento RNA-Seq che hanno per oggetto homo sapiens e trattano una malattia professionale specifica magari derivata dall'inalazione di una particolare sostanza (Figura 6).

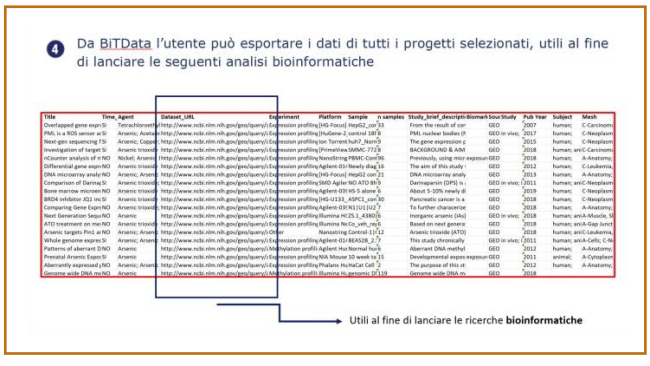

Figura 6. Cosa può fare l'utente? Oltre alla consultazione degli articoli scientifici e dei dati genomici scaricati dai repository è possibile scaricare il file riepilogativo dei progetti di interesse in cui sono presenti i link dei dataset_url necessari per le analisi bioinformatiche.

Elaborazione di dati attraverso specifico tool della dashboard della piattaforma di IA

Il ricercatore Inail, che ne ha le credenziali, può sia accedere direttamente alla piattaforma di IA per fare delle analisi puntuali sugli esperimenti che ha raccolto prima che vengano salvati sulla banca dati Bit Data; sia elaborare grafici specifici di interesse per la ricerca.
In questo fact sheet presentiamo alcuni esempi di elaborazioni di dati che riteniamo interessanti.
Inizialmente abbiamo voluto indagare quali fossero i metodi sperimentali utilizzati per studiare i contaminanti occupazionali causa di patologie respiratorie. Dal grafico rappresentato in figura 7 emerge che le sostanze come metalli pesanti ed idrocarburi sono state prevalentemente studiate tramite tecniche di microarray, utilizzate enormemente nel decennio precedente. Per quanto riguarda invece, ad esempio, gli studi su SARS-CoV-2, relativamente recenti si evince come ormai sia preponderante l'utilizzo di tecniche di profiling attraverso high troughput expression.

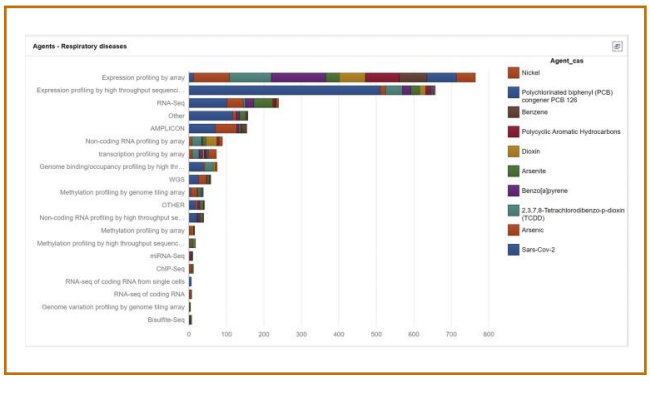

Figura 7. Grafico della frequenza dei dati molecolari depositati degli agenti causa di patologie respiratorie in relazione ai metodi sperimentali principalmente utilizzati.

Analizzando il grafico rappresentato in figura 8 quali sostanze siano state maggiormente studiate nel tempo, dal 2000 al 2023, ad esempio si può notare come indubbiamente nel triennio precedente nella totalità degli studi molecolari condotti sia stata preponderante la presenza di SARS-CoV-2.

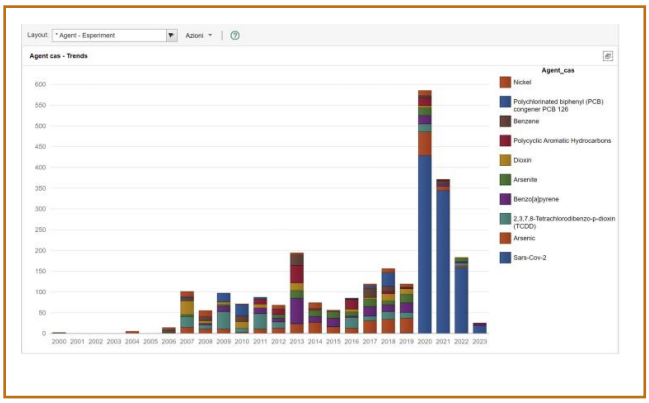

Figura 8. Grafico della frequenza dei dati molecolari depositati degli agenti causa di patologie respiratorie negli anni 2000-2023.

Nel grafico rappresentato in figura 9 abbiamo navigato i dati molecolari raccolti dai repository e filtrato solo quelli relativi all'esposizione occupazionale e li abbiamo messi in relazione ai metodi sperimentali principalmente utilizzati.

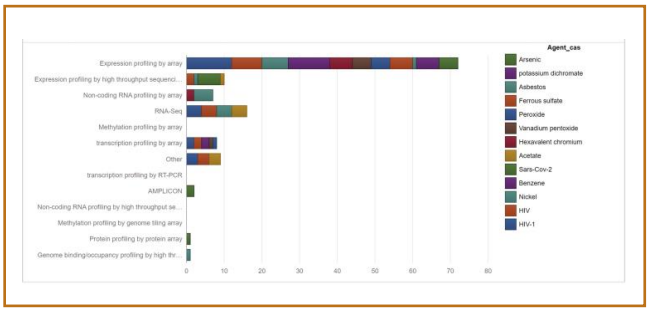

Figura 9. Grafico della frequenza dei dati molecolari depositati come relativi all'esposizione occupazionale in relazione ai metodi sperimentali principalmente utilizzati.

Il grafico a linee rette relativo agli studi su diversi agenti occupazionali e condotti nelle diverse annualità riflette anche l'attenzione dei ricercatori a condurre studi per indagare determinate sostanze in relazione all'emergenze di casi di cronaca, di provvedimenti legislativi e/o che riflettono il sentiment dell'opinione pubblica (Figura 10).

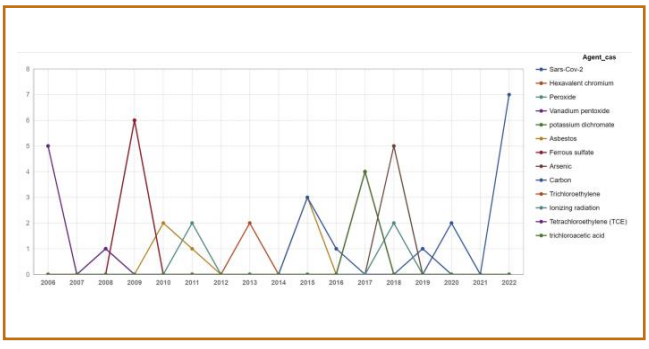

Figura 10. Grafico della frequenza dei dati molecolari depositati degli agenti occupazionali negli anni 2006-2022.

Andando a navigare i dati attraverso uno specifico tool della dashboard che permette di ottenere una Word Cloud di Correlazione tra i parametri da analizzare, si evince che gli studi molecolari finora condotti per gli agenti occupazionali sono principalmente studi *in vitro* e su cellule di carcinoma, carcinoma polmonare non a piccole cellule (agenti respiratori) e di neuroblastoma. Questo dato riflette infatti la difficoltà di ottenere una numerosità statisticamente rilevante di biopsie (studi *in vivo*) significative per gli studi di esposizione, per cui gli studi molecolari sono principalmente costituiti da studi in vitro. Inoltre, gli studi molecolari maggiormente documentati relativi a xenobiotici occupazionali sono quelli condotti per HIV-1, Nickel, Benzene, SARS-CoV-2, Cromo esavalente, Arsenico (Figura 11).

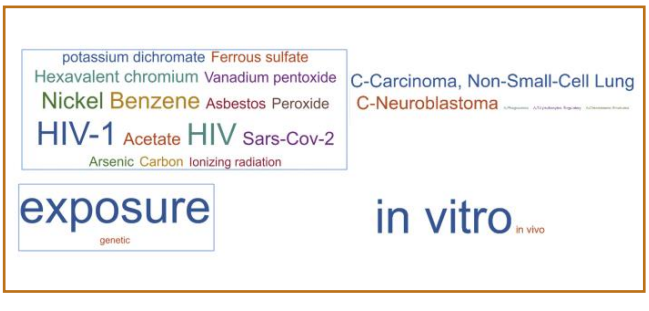

Figura 11. Word Cloud di Correlazione tra i parametri analizzati.

RIFERIMENTI BIBLIOGRAFICI

[1] Satam H., Joshi K., Mangrolia U., Waghoo S., Zaidi G., Rawool S., Thakare R.P., Banday S., Mishra A.K., Das G., Malonia S.K. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology. 2023; 12(7):997. https://doi.org/10.3390/biology12070997

[2] Sturchio E., Berardinelli M.G., Boccia P., Zanellato M., Gioiosa S., 2020: MicroRNAs diagnostic and prognostic value as pre-dictive markers for malignant mesothelioma, Archives of Environmental & Occupational Health, DOI: 10.1080/19338244.2020.1747966.

[3] Gioiosa S., Berardinelli M.G., Paradisi A., Boccia P., Zanellato M., Ceruti F., Sturchio E. "Sviluppo della banca dati molecolare Inail (BiTdata) come utile strumento per studi di esposizione occupazionale, Rivista Degli Infortuni E Delle Malattie Profes-sionali Fascicolo N. 3/2018, 487 502.

[4] Sturchio E., Zanellato M., Boccia P., Meconi C., Gioiosa S. Pos-sibile ruolo di microRNA come biomarcatori di esposizione ad amianto e del mesotelioma pleurico maligno. In Minoia C., Comba P. Amianto, un fantasma del passato o una storia infi-nita? Cermenate (CO), New Press Edizioni, 2018 2 volumi, 980 pagine. ISBN/EAN 9788893560382

[5] Sturchio E., Colombo T., Boccia P., Carucci N., Meconi C., Mi-noia C., Macino G.: Arsenic exposure triggers a shift in microRNA expression Science of Total Environment, 472 (2014) 672-680. doi: 10.1016/j.scitotenv.2013.11.092.

ISBN 978-88-7484-831-7

Stampato dalla Tipolitografia Inail di Milano • Edizione 2023 **Autori: E. Sturchio, P. Boccia, M. Zanellato (Inail Dit), S. Gioiosa (Cineca), L. Di Piramo (IBM Italia) • Grafica ed Editing: Cogliani D.G., Ceruti F.**

Inail - Dipartimento innovazioni tecnologiche e sicurezza degli impianti, prodotti e insediamenti antropici - Direttore: Corrado Delle Site
Via Roberto Ferruzzi, 38/40 - 00143 Roma
Tel. 06 548 76 400 • Fax 06 548 76 413 • dit@inail.it